SUBMITTED TO APJ
Preprint typeset using LaTeX style emulateapj v. 10/09/06

SUPERMASSIVE BLACK HOLES IN THE HIERARCHICAL UNIVERSE: A GENERAL FRAMEWORK AND OBSERVATIONAL TESTS

YUE SHEN
PRINCETON UNIVERSITY OBSERVATORY, PRINCETON, NJ 08544.

Submitted to ApJ

ABSTRACT

We present a simple framework for the growth and evolution of supermassive black holes (SMBHs) in the hierarchical structure formation paradigm, adopting the general idea that quasar activity is triggered in major mergers. In our model, black hole accretion is triggered during major mergers (mass ratio $\gtrsim 0.3$) between host dark matter halos. The successive evolution of quasar luminosities follows a universal light curve form, during which the growth of the SMBH is modeled self-consistently: an initial exponential growth at constant Eddington ratio of order unity until it reaches the peak luminosity, followed by a power-law decay. Assuming that the peak luminosity correlates with the post-merger halo mass, we convolve the light curve with the triggering rate of quasar activity to predict the quasar luminosity function (LF). Our model reproduces the observed LF at $0.5 < z < 4.5$ for the full luminosity ranges probed by current optical and X-ray surveys. At $z < 0.5$, our model underestimates the LF at $L_{\rm bol} < 10^{45}\,{\rm ergs}^{-1}$, allowing room for AGN activity triggered by secular processes instead of major mergers. At $z > 4.5$, in order to reproduce the observed quasar abundance, the typical quasar hosts must shift to lower mass halos, and/or minor mergers can also trigger quasar activity. Our model reproduces both the observed redshift evolution and luminosity dependence of the linear bias of quasar/AGN clustering. Due to the scatter between instantaneous luminosity and halo mass, quasar/AGN clustering weakly depends on luminosity at low to intermediate luminosities; but the linear bias rises rapidly with luminosity at the high luminosity end and at high redshift. In our model, the Eddington ratio distribution is roughly log-normal, which broadens and shifts to lower mean values from high luminosity quasars ($L_{\rm bol} \gtrsim 10^{46}\,{\rm ergs}^{-1}$) to low luminosity AGNs ($L_{\rm bol} \lesssim 10^{45}\,{\rm ergs}^{-1}$), in good agreement with observations. The model predicts that the vast majority of $\gtrsim 10^{8.5}\,M_{\odot}$ SMBHs were already in place by $z = 1$, and $\lesssim 50\%$ of them were in place by $z = 2$; but the less massive ($\lesssim 10^{7}\,M_{\odot}$) SMBHs were assembled more recently, likely more through secular processes than by major mergers – in accordance with the downsizing picture of SMBH assembly since the peak of bright quasar activity around $z \sim 2-3$.

Subject headings: black hole physics – galaxies: active – cosmology: observations – large-scale structure of universe – quasars: general – surveys

1. INTRODUCTION

The cosmic assembly and evolution of supermassive black holes (SMBHs) is a central topic in modern cosmology: in the local universe, SMBHs reside in almost every bulge dominant galaxy (e.g., Kormendy & Richstone 1995; Richstone et al. 1998); and they likely played important roles during their coevolution with galaxy bulges (e.g., Silk & Rees 1998; Wyithe & Loeb 2002, 2003; King 2003; Di Matteo et al. 2005; Croton et al. 2006), as inferred from observed scaling relations between bulge properties and the mass of the central BH (e.g., Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Graham et al. 2001; Tremaine et al. 2002; Marconi & Hunt 2003). It is also generally accepted that the local dormant SMBH population was largely assembled via gas accretion during a luminous QSO phase[1] (e.g., Salpeter 1964; Zel'dovich & Novikov 1964; Lynden-Bell 1969; Soltan 1982; Small & Blandford 1992; Salucci et al. 1999; Yu & Tremaine 2002; Shankar et al. 2004; Marconi et al. 2004). The statistics of the local dormant SMBHs and distant active QSOs therefore provide clues to the build-up of the local BH density across cosmic time, and have implications for the hierarchical structure formation paradigm, as well as for the interactions between accreting SMBHs and galaxy bulges.

One of the leading hypotheses for the QSO triggering mechanism is galaxy mergers (e.g., Hernquist 1989; Carlberg 1990; Kauffmann & Haehnelt 2000; Hopkins et al. 2008, and references therein). In the hierarchical CDM paradigm, small structures merge to form large structures, and the merger rate of dark matter halos peaks at early times, in broad consistency with the observed peak of bright quasar activities. Gas-rich major mergers (i.e., those with comparable mass ratios) between galaxies provide an efficient way to channel a large amount of gas into the central region to trigger starbursts and possibly feed rapid black hole growth. Observationally this merger hypothesis is supported by the ULIRG-quasar connection (e.g., Sanders & Mirabel 1996; Canalizo & Stockton 2001), the signature of recent mergers in quasar hosts (e.g., Bennert et al. 2008), and the small-scale overdensities of galaxies around luminous quasars or quasar binaries (e.g., Fisher et al. 1996; Bahcall et al. 1997; Serber et al. 2006; Hennawi et al. 2006; Myers et al. 2008; Strand et al. 2008). These observations do not necessarily prove that mergers are directly responsible for triggering QSO activity, but they at least suggest that QSO activity is coincident with mergers in many cases.

The last decade has seen a number of models for QSO evolution based on the merger hypothesis (e.g., Haiman & Loeb

[1] In this paper we use the term "QSO" to refer to all actively accreting SMBHs. We adopt the convention that $L_{\rm bol} = 10^{45}\,{\rm ergs}^{-1}$ is the dividing line between quasars and AGNs. This division is slightly lower than the traditional Seyfert/quasar division (Schmidt & Green 1983) of $M_B = -23$ or $L_{\rm bol} \sim 10^{12}\,L_{\odot}$.

1998; Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003; Volonteri et al. 2003; Hopkins et al. 2008), which agree with the bulk of QSO observations reasonably well. Motivated by the success of these merger-based QSO models, we revisit this problem in this paper with improved observational data on SMBHs and QSOs from dedicated large surveys, and with updated knowledge of the merger rate as inferred from recent numerical simulations. The present work is different from previous studies in many aspects in both methodology and the observational tests used. Our goal is to check if a simple merger-based cosmological QSO framework can reproduce all the observed statistics of SMBHs (both active and dormant), and to put constraints on the physical properties of SMBH growth. Our model is observationally motivated, therefore we do not restrict ourselves to theoretical predictions of SMBH/QSO properties from either cosmological or merger event simulations.

The paper is organized as follows. In §1.1 we review some aspects of local SMBH demographics and QSO luminosity function; in §1.2 we briefly review the current status of quasar clustering observations; §§1.3 and 1.4 discuss the halo and subhalo merger rate from numerical simulations. Our model formulism is presented in §2 and we present our fiducial model and compare with observations in §3. We discuss additional aspects of our model in §4 and conclude in §5.

We use *friends-of-friends (fof)* mass with a link length $b = 0.2$ to define the mass of a halo (roughly corresponding to spherical overdensity mass with $\Delta \sim 200$ times the mean *matter* density), since the fitting formulae for the halo mass function are nearly universal with this mass definition (Sheth et al. 2001; Jenkins et al. 2001; Warren et al. 2006; Tinker et al. 2008). For simplicity we neglect the slight difference between *fof* mass and virial mass (Appendix A; and see White 2002, for discussions on different mass definitions). Throughout the paper, L always refers to the bolometric luminosity, and we use subscripts $_B$ or $_X$ to denote B-band or X-ray luminosity when needed. We adopt a flat ΛCDM cosmology with $\Omega_0 = 0.26$, $\Omega_\Lambda = 0.74$, $\Omega_b = 0.044$, $h = 0.7$, $\sigma_8 = 0.78$, $n_s = 0.95$. We use the Eisenstein & Hu (1999) transfer function to compute the linear power spectrum, and use the fitting formulae for halo abundance from Sheth & Tormen (1999) and for halo linear bias from Sheth et al. (2001) based on the ellipsoidal collapse model and tested against numerical simulations.

1.1. *Supermassive Black Hole Demographics*

In the local universe, the dormant BH mass function (BHMF) can be estimated by combining scaling relations between BH mass and galaxy bulge properties, such as the $M_\bullet - \sigma$ relation (e.g., Gebhardt et al. 2000; Ferrarese & Merritt 2000; Tremaine et al. 2002) or the $M_\bullet - L_{\rm sph}(M_{\rm sph})$ relation (e.g., Magorrian et al. 1998; Marconi & Hunt 2003), with bulge luminosity or stellar mass/velocity dispersion functions (e.g., Salucci et al. 1999; Yu & Tremaine 2002; Marconi et al. 2004; Shankar et al. 2004; McLure & Dunlop 2004; Yu & Lu 2004, 2008; Tundo et al. 2007; Shankar et al. 2009b). Although some uncertainties exist on the usage of these scaling relations at both the high and low BH mass end (e.g., Lauer et al. 2007; Tundo et al. 2007; Hu 2008; Greene et al. 2008; Graham 2008; Graham & Li 2009), the total BH mass density is estimated to be $\rho_\bullet \approx 4\times10^5\,M_\odot{\rm Mpc}^{-3}$ with an uncertainty of a factor ~ 1.5 (e.g. Yu & Lu 2008; Shankar et al. 2009b) – but the shape of the local BHMF is more uncertain (see the discussion in Shankar et al. 2009b).

It is now generally accepted that local SMBHs were once luminous QSOs (Salpeter 1964; Zel'dovich & Novikov 1964; Lynden-Bell 1969). An elegant argument tying the active QSO population to the local dormant SMBH population was proposed by Soltan (1982): if SMBHs grow mainly through a luminous QSO phase, then the accreted luminosity density of QSOs to $z = 0$ should equal the local BH mass density:

$$\rho_{\bullet,\rm acc} = \int_0^\infty \frac{dt}{dz}dz \int_0^\infty \frac{(1-\epsilon)L}{\epsilon c^2}\Phi(L,z)dL \approx \rho_\bullet \ , \qquad (1)$$

where $\Phi(L,z)$ is the bolometric luminosity function (LF) per L interval. Given the observed QSO luminosity function, a reasonably good match between $\rho_{\bullet,\rm acc}$ and $\rho_\bullet$ can be achieved if the average radiative efficiency $\epsilon \sim 0.1$ (e.g., Yu & Tremaine 2002; Shankar et al. 2004; Marconi et al. 2004; Hopkins et al. 2007; Shankar et al. 2009b). The exact value of ϵ, however, is subject to some uncertainties from the luminosity function and local BH mass density determination.

Assuming that SMBH growth is through gas accretion, an extended version of the Soltan argument can be derived using the continuity equation[2] (cf. Small & Blandford 1992):

$$\frac{\partial n(M_\bullet,t)}{\partial t} + \frac{\partial[n(M_\bullet,t)\langle\dot{M}_\bullet\rangle]}{\partial M_\bullet} = 0 \ , \qquad (2)$$

where $n(M_\bullet,t)$ is the BH mass function per $dM_\bullet$, and $\langle\dot{M}_\bullet\rangle$ is the mean accretion rate of BHs with mass $(M_\bullet, M_\bullet + dM_\bullet)$ averaged over both active and inactive BHs. Given the luminosity function, and a model connecting luminosity to BH mass, one can derive the BH mass function at all times by integrating the continuity equation (e.g., Marconi et al. 2004; Merloni 2004; Shankar et al. 2009b).

Since the local BHMF and the QSO luminosity function together determine the assembly history of the cosmic SMBH population, any cosmological model of AGN/quasars must first reproduce the observed luminosity function (e.g., Haiman & Loeb 1998; Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003; Volonteri et al. 2003; Lapi et al. 2006; Hopkins et al. 2008; Shankar et al. 2009b). This is also one of the central themes of this paper.

There has been great progress in the measurements of the QSO luminosity function over wide redshift and luminosity ranges for the past decade, mostly in the optical band (e.g., Fan et al. 2001, 2004; Wolf et al. 2003; Croom et al. 2004; Richards et al. 2005, 2006; Jiang et al. 2006; Fontanot et al. 2007), and the soft/hard X-ray band (e.g., Ueda et al. 2003; Hasinger et al. 2005; Silverman et al. 2005; Barger et al. 2005). While wide-field optical surveys provide the best constraints on the bright end of the QSO luminosity functions, deep X-ray surveys can probe the obscured faint end AGN population. Although it has been known for a while that the spatial density of optically-selected bright quasars peaks at redshift $z \sim 2-3$, the spatial density of fainter AGNs selected in the X-ray seems to peak at lower redshift (e.g., Steffen et al. 2003; Ueda et al. 2003; Hasinger et al. 2005), a trend now confirmed in other observational bands as well (e.g., Bongiorno et al. 2007; Hopkins et al. 2007, and references therein) – the manifestation of the so-called *downsizing* of the cosmic SMBH assembly.

[2] Note that here the source term, i.e., the creation of a BH with mass $M_\bullet$ by mechanisms other than gas accretion, is generally neglected (e.g., Small & Blandford 1992; Merloni 2004; Shankar et al. 2009b). One possible such mechanism is BH coalescence, which modifies the shape of the BH mass function but does not change the total BH mass density in the classical case (i.e., neglecting mass loss from gravitational radiation). Dry mergers at low redshift may act as a surrogate of shaping the BH mass function.

Hopkins et al. (2007) compiled QSO luminosity function data from surveys in various bands (optical, X-ray, IR, etc). Assuming a general AGN spectral energy distribution (SED) and a column density distribution for obscuration, Hopkins et al. (2007) were able to fit a universal bolometric luminosity function based on these data. We adopt their compiled bolometric LF data in our modeling. The advantage of using the bolometric LF data is that both unobscured and obscured SMBH growth are counted in the model; but we still suffer from the systematics of bolometric corrections. The nominal statistical/systematic uncertainty level of the bolometric LF is $\sim 20-30\%$ (cf., Shankar et al. 2009b).

1.2. *Quasar Clustering*

A new ingredient in SMBH studies, due to dedicated large-scale optical surveys such as the 2QZ (Croom et al. 2004) and SDSS (York et al. 2000), is quasar clustering. While quasar clustering studies can be traced back to more than two decades ago (e.g., Shaver 1984), statistically significant results only came very recently (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Myers et al. 2006, 2007a,b; Shen et al. 2007; da Ângela et al. 2008; Padmanabhan et al. 2008; Shen et al. 2008b, 2009; Ross et al. 2009).

Within the biased halo clustering picture (e.g., Kaiser 1984; Bardeen et al. 1986; Mo & White 1996), the observed QSO two-point correlation function implies that quasars live in massive dark matter halos and are biased tracers of the underlying dark matter. For optically luminous quasars ($L_{\rm bol} \gtrsim 10^{45}\,{\rm ergs}^{-1}$), the host halo mass inferred from clustering analysis is a few times $10^{12}\,h^{-1}M_{\odot}$. Thus quasar clustering provides independent constraints on how SMBHs occupy dark matter halos, where the abundance and evolution of the latter population can be well studied in analytical theories and numerical simulations (e.g., Press & Schechter 1974; Bond et al. 1991; Lacey & Cole 1993; Mo & White 1996; Sheth & Tormen 1999; Sheth et al. 2001; Springel et al. 2005b). By comparing the relative abundance of quasars and their host halos, one can constrain the average quasar duty cycles or lifetimes (e.g., Cole & Kaiser 1989; Martini & Weinberg 2001; Haiman & Hui 2001) to be $\lesssim 10^8$ yr, which means bright quasars are short lived.

More useful constraints on quasar models come from studies of quasar clustering as a function of redshift and luminosity. The redshift evolution of quasar clustering constrains the evolution of duty cycles of active accretion, while the luminosity dependence of quasar clustering puts constraints on quasar light curves (LC). Successful cosmological quasar models therefore not only need to account for quasar abundances (LF), but also need to explain quasar clustering properties, both as function of redshift and luminosity. We will use quasar clustering observations as additional tests on our quasar models, as has been done in recent studies (e.g., Hopkins et al. 2008; Shankar et al. 2009a; Thacker et al. 2009; Bonoli et al. 2009; Croton 2009).

The luminosity dependence of quasar bias can be modeled as follows. Assume at redshift z the probability distribution of host halo mass given bolometric luminosity L is $p(M|L,z)$, then the halo averaged bias factor is:

$$b_L(L,z) = \int b_M(M,z)p(M|L,z)dM\ , \qquad (3)$$

where $b_M(M,z)$ is the linear bias factor for halos with mass M at redshift z. The derivation of the probability distribution $p(M|L,z) \equiv dP(M|L,z)/dM$ is described in later sections [Eqn. (20)].

1.3. *Dark Matter Halo Mergers*

In the hierarchical universe small density perturbations grow under gravitational forces and collapse to form virialized halos (mostly consisting of dark matter). Smaller halos coalesce and merge into larger ones. Early work on the merger history of dark matter halos followed the extended Press-Schechter (EPS) theory (e.g., Press & Schechter 1974; Bond et al. 1991; Lacey & Cole 1993), which is based on the spherical collapse model (Gunn & Gott 1972) with a constant barrier for the collapse threshold (the critical linear overdensity $\delta_c \approx 1.68$ is independent on mass, albeit it depends slightly on cosmology in the ΛCDM universe). Although the (unconditional) halo mass function $n(M,z)$ predicted by the EPS theory agrees with numerical N-body simulations reasonably well, it is well known that it overpredicts (underpredicts) the halo abundance at the low (high) mass end (Sheth & Tormen 1999, and references therein). Motivated by the fact that halo collapses are generally triaxial, Sheth & Tormen (1999), based on earlier work by Bond & Myers (1996), proposed the ellipsoidal collapse model with a moving barrier where the collapse threshold also depends on mass. By imposing this mass dependence on collapse barrier, the abundance of small halos is suppressed and fitting formulae for the (unconditional) halo mass function are obtained, which agree with N-body simulations much better than the spherical EPS predictions (e.g., Sheth & Tormen 1999; Sheth et al. 2001; Jenkins et al. 2001; Warren et al. 2006; Tinker et al. 2008).

In addition to the *unconditional* halo mass function, the *conditional* mass function $n(M_1,z_1|M_0,z_0)$ gives the mass spectrum of progenitor halos at an earlier redshift z_1 of a descendant halo M_0 at redshift z_0. This conditional mass function thus contains information about the merger history of individual halos and can be used to generate halo merger trees. A simple analytical form for $n(M_1,z_1|M_0,z_0)$ can be obtained in the spherical EPS framework (e.g., Lacey & Cole 1993); for the ellipsoidal collapse model, an exact but computationally consuming solution of the conditional mass function can be obtained by solving the integral equation proposed by Zhang & Hui (2006). Recently, Zhang et al. (2008) derived analytical formulae for $n(M_1,z_1|M_0,z_0)$ in the limit of small look-back times for the ellipsoidal collapse model.

Alternatively, the halo merger rate can be retrieved directly from large volume, high resolution cosmological N-body simulations, which bypasses the inconsistencies between the spherical EPS theory and numerical simulations. Using the product of the Millennium Simulation (Springel et al. 2005b), Fakhouri & Ma (2008) quantified the mean halo merger rate per halo for a wide range of descendant (at $z = 0$) halo mass $10^{12} \lesssim M_0 \lesssim 10^{15}\,M_{\odot}$, progenitor mass ratio $10^{-3} \lesssim \xi \leq 1$ and redshift $0 \leq z \lesssim 6$, and they provided an almost universal fitting function for the mean halo merger rate to an accuracy $\lesssim 20\%$ within the numerical simulation results:

$$\frac{B(M_0,\xi,z)}{n(M_0,z)} = 0.0289\left(\frac{M_0}{1.2\times10^{12}M_{\odot}}\right)^{0.083}\xi^{-2.01}\times \exp\left[\left(\frac{\xi}{0.098}\right)^{0.409}\right]\left(\frac{d\delta_{c,z}}{dz}\right)^{0.371}. \qquad (4)$$

Here $B(M_0,\xi,z)$ is the instantaneous merger rate at redshift z, for halos with mass (M_0, M_0+dM_0) and with progeni-

tor mass ratio in the range $(\xi, \xi+d\xi)$ [$B(M_0,\xi,z)$ is in units of number of mergers $\times \mathrm{Mpc}^{-3} M_\odot^{-1} dz^{-1} d\xi^{-1}$], $n(M_0,z)$ is the halo mass function (in units of $\mathrm{Mpc}^{-3} {M_\odot}^{-1}$), $\xi \equiv M_2/M_1 \leq 1$ where $M_1 \geq M_2$ are the masses of the two progenitors, and $\delta_{c,z} = \delta_c/D(z)$ is the linear density threshold for spherical collapse with $D(z)$ the linear growth factor. We adopt Eqn. (4) to estimate the halo merger rate in our modeling. The mass definition used here is *friends-of-friends* mass $M_{\rm fof}$ with a link length $b = 0.2$. For the range of mass ratios we are interested in (i.e., major mergers), the halo merger rate derived from the spherical EPS model can overpredict the merger rate by up to a factor of ~ 2 (Fakhouri & Ma 2008).

1.4. *Galaxy (subhalo) Mergers*

Galaxies reside in the central region of dark matter halos where the potential well is deep and gas can cool to form stars. When a secondary halo merges with a host halo, it (along with its central galaxy) becomes a satellite within the virial radius of the host halo. It will take a dynamical friction time for the subhalo to sink to the center of the primary halo, where the two galaxies merge. The subsequent galaxy (subhalo) merger rate is therefore different from the halo merger rate discussed in §1.3, and a full treatment with all the dynamics (tidal stripping and gravitational shocking of subhalos) and baryonic physics is rather complicated.

The simplest argument for the galaxy (subhalo) merger timescale is given by dynamical friction (Chandrasekhar 1943; Binney & Tremaine 1987):

$$\tau_{\rm df} \approx \frac{f_{\rm df}\Theta_{\rm orb}}{\ln\Lambda}\frac{M_{\rm host}}{M_{\rm sat}}\tau_{\rm dyn}\ , \qquad (5)$$

where $M_{\rm host}$ and $M_{\rm sat}$ are the masses for the host and satellite halos respectively, $\ln\Lambda \approx \ln(M_{\rm host}/M_{\rm sat})$ is the Coulomb logarithm, $\Theta_{\rm orb}$ is a function of the orbital energy and angular momentum of the satellite, $f_{\rm df}$ is an adjustable parameter and $\tau_{\rm dyn} \equiv r/V_c(r)$ is the halo dynamical timescale, usually estimated at the halo virial radius $r_{\rm vir}$. This formula (5) is valid at the small satellite mass limit $M_{\rm host}/M_{\rm sat} \gg 1$; although it is also used in cases of $M_{\rm host} \gtrsim M_{\rm sat}$ in many semi-analytical models (SAMs) with the replacement of $\ln\Lambda = (1/2)\ln[1+(M_{\rm host}/M_{\rm sat})^2]$ (i.e., the original definition of the Coulomb logarithm) or $\ln\Lambda = \ln(1+M_{\rm host}/M_{\rm sat})$. However, in recent years deviations from the predictions by Eqn. (5) in numerical simulations have been reported for both the $M_{\rm sat} \ll M_{\rm host}$ and the $M_{\rm sat} \lesssim M_{\rm host}$ regimes (e.g., Taffoni et al. 2003; Monaco et al. 2007; Boylan-Kolchin et al. 2008; Jiang et al. 2008). In particular even in the regime $M_{\rm sat}/M_{\rm host} \ll 1$ where the original Chandrasekhar formula is supposed to work, Eqn. (5) substantially underestimates the merger timescale from simulations by a factor of a few (getting worse at lower mass ratios). This is because Eqn. (5) is derived by treating the satellite as a rigid body, while in reality the satellite halo loses mass via tidal stripping[3] and hence the duration of dynamical friction is greatly extended.

Given the limitations of analytical treatments, we therefore retreat to numerical simulation results. Fitting formulae for the galaxy merger timescales within merged halos have been provided by several groups (Taffoni et al. 2003; Monaco et al. 2007; Boylan-Kolchin et al. 2008; Jiang et al. 2008), and the merger rate of subhalos has also been directly measured from simulations (Wetzel et al. 2009; Stewart et al. 2008).

[3] The galaxy associated with the subhalo, on the other hand, does not suffer from mass stripping since it sits in the core region of the subhalo.

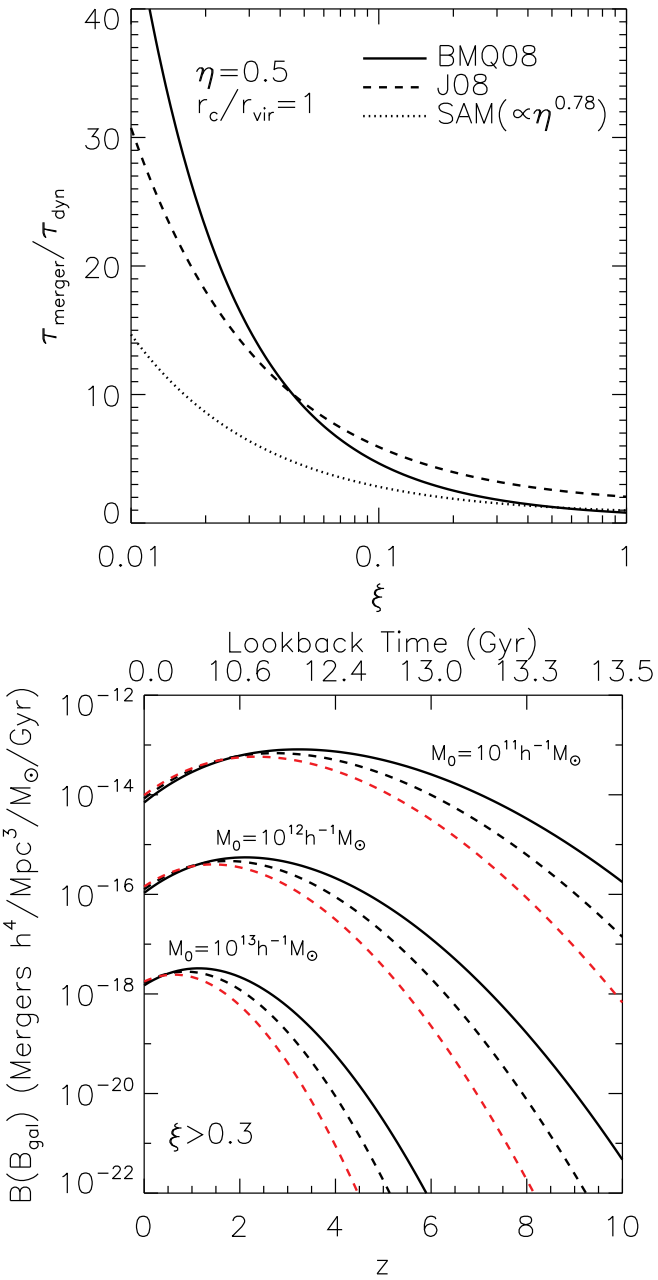

FIG. 1.— Merger timescales and (sub)halo merger rates. *Upper*: comparison of the subhalo merger timescale from three dynamical friction prescriptions. *Bottom*: the redshift evolution of the halo merger rate (black solid lines), and the subhalo merger rate using the Boylan-Kolchin et al. (2008) (black dashed lines), and Jiang et al. (2008) (red dashed lines) prescriptions for the merger timescale, for three (post)merger halo masses.

The fitting formula for $\tau_{\rm merger}$ in Jiang et al. (2008), who used hydro/N-body simulations, has the following form:

$$\tau_{\rm merger}(\xi,z) = \frac{0.94\eta^{0.6}+0.6}{2\times 0.43}\frac{1}{\xi\ln[1+(1/\xi)]}\frac{r_{\rm vir}}{V_c}\ , \qquad (6)$$

where η is the circularity parameter [$\eta = (1-e^2)^{1/2}$, and e is the orbit eccentricity]. In deriving this equation Jiang et al. have removed the energy dependence of the orbit, i.e., they fix $r_c = r_{\rm vir}$ where r_c is the circular orbit that has the same energy as the satellite's orbit. In their simulation the ratio of $r_c/r_{\rm vir}$ ranges from $\sim 0.6-1.5$ and has a median value ~ 1. The distribution of circularity η in their simulation (see their eqn. 7) has a median value ~ 0.5, so in what follows

we take $\eta = 0.5$. The redshift dependence of $\tau_{\rm merger}$ is thus the same as that of the halo dynamical time $\tau_{\rm dyn} = r_{\rm vir}/V_c \propto [\Delta_{\rm vir}(z)]^{-1/2}(1+z)^{-3/2}$ (see Appendix A).

Alternatively, the fitting formula of $\tau_{\rm merger}$ in Boylan-Kolchin et al. (2008) is given by:

$$\tau_{\rm merger}(\xi,z) = 0.216\frac{\exp(1.9\eta)}{\xi^{1.3}\ln[1+(1/\xi)]}\frac{r_c(E)}{r_{\rm vir}}\tau_{\rm dyn}\ . \qquad (7)$$

Combining the halo merger rate (4) and galaxy (subhalo) merger timescale (6) or (7) we derive the instantaneous merger rate of galaxies within a merged halo of mass M_0 for progenitor mass ratio ξ and at redshift z:

$$B_{\rm gal}(M_0,\xi,z) = B[M_0,\xi,z_{\rm e}(z,\xi)]\frac{dz_{\rm e}}{dz}\ , \qquad (8)$$

where $z_{\rm e}(z,\xi)$ is a function of (z,ξ) and is determined by:

$$t_{\rm age}(z) - t_{\rm age}(z_{\rm e}) = \tau_{\rm merger}(\xi,z_{\rm e})\ , \qquad (9)$$

where $t_{\rm age}(z)$ is the cosmic time at redshift z. Again, here $B_{\rm gal}(M_0,\xi,z)$ is the number of mergers per volume per halo mass per mass ratio per redshift. We find that $dz_{\rm e}/dz$ is almost constant *at all redshifts* for $\xi = 0.1-1$ and monotonically decreases as ξ increases. This constant can be approximated by its asymptotic value at large z when $\Omega(z) \to 1$:

$$\frac{dz_{\rm e}}{dz} \approx \left\{1+0.22\left[\xi\ln(1+1/\xi)\right]^{-1}\right\}^{2/3}\ , \qquad (10)$$

for the fitting formula of $\tau_{\rm merger}$ in Jiang et al. (2008), or

$$\frac{dz_{\rm e}}{dz} \approx \left\{1+0.09\left[\xi^{1.3}\ln(1+1/\xi)\right]^{-1}\right\}^{2/3}\ , \qquad (11)$$

for the fitting formula of $\tau_{\rm merger}$ in Boylan-Kolchin et al. (2008).

Unfortunately, current studies on the galaxy merger timescale have not converged yet, and different groups do report similar but quantitatively different results, at least partly caused by the different definitions and setups in their numerical simulations (e.g., hydro/N-body versus pure N-body simulations, halo finding algorithms and definition of mergers, etc). These fitting formulae are generally good within a factor of ~ 2 uncertainty. This uncertainty in the galaxy merger timescale within DM halos will lead to quite substantial differences in the galaxy merger rate at high redshift. To see this, we show in Fig. 1 the comparison of different fitting formulae for $\tau_{\rm merger}$ and their consequences. In the upper panel of Fig. 1 we show the ratio of $\tau_{\rm merger}/\tau_{\rm dyn}$ for the most likely orbit with circularity $\eta = 0.5$ and $r_c = r_{\rm vir}$ using the fitting formula in Boylan-Kolchin et al. (2008) (solid line), Jiang et al. (2008) (dashed line), and a commonly-used SAM prescription: $\tau_{\rm merger} = 1.16\eta^{0.78}[\xi\ln(1+1/\xi)]^{-1}r_c/r_{\rm vir}$ (dotted line). For the major merger regime $\xi \gtrsim 0.3$ the fitting formula in Boylan-Kolchin et al. (2008) agrees with the SAM prescription well, and they both approach the dynamical time at the high mass ratio end. However, the fitting formula in Jiang et al. (2008) yields a factor of ~ 2 longer than dynamical time at the high mass ratio end.

In the lower panel of Fig. 1 we show the halo and galaxy major merger rates *per unit time* (integrated over $\xi > 0.3$), as function of redshift. The black solid lines show the halo major merger rate, the black dashed lines show the galaxy merger rate using the fitting formula of $\tau_{\rm merger}$ from Boylan-Kolchin et al. (2008), and the red dashed lines show the galaxy merger rate adopting the fitting formula from Jiang et al. (2008). In general the galaxy merger rates fall below the halo merger rate at high redshift and take over at lower redshift. At redshift $z > 4$, the galaxy merger rate using the Jiang et al. (2008) formula is lower by almost an order of magnitude than that using the Boylan-Kolchin et al. (2008) formula (as well as the SAM prescription, since it agrees with eqn. 7 for $\xi > 0.3$), which makes it difficult to produce the quasar abundance at high redshift (see later sections).

It is worth noting that although the merger rate peaks at some redshift, this trend is *hierarchical* such that it peaks at later time for more massive halos. This is somewhat in contradiction to the *downsizing* of the QSO luminosity function. This apparent discrepancy is likely caused by the fact that at low reshift, it becomes progressively more difficult for the black hole to accrete efficiently in massive halos because of the global deficit of cold gas due to previous mergers experienced by massive halos and/or possible feedback quenches (e.g., Kauffmann & Haehnelt 2000; Croton et al. 2006). We will treat this fraction of QSO-triggering merger event as function of redshift and halo mass explicitly in our modeling.

If we choose to normalize the total merger rate B ($B_{\rm gal}$) by $n(M_0,z)$, the abundance of descendant halos with mass M_0, at the same redshift z for halo mergers and galaxy mergers, then $B_{\rm gal}/n$ falls below B/n at high redshift, then catches up and exceeds B/n a little, and evolves more or less parallel to B/n afterwards. Therefore the evolution in $B_{\rm gal}/n$ is shallower than that in B/n, consistent with the findings by Wetzel et al. (2009) once the different definitions of B/n and halo mass are taken into account.

To summarize §1.3 and §1.4, we have compared the halo merger rate and galaxy merger rate as functions of redshift, (post)merger halo mass and mass ratio with different prescriptions for the dynamical friction timescale. It is, however, unclear on which rate we should link to the QSO triggering rate. It is true that it will take a dynamical friction time for the two galaxies to merge after their host halo merged. But the black hole accretion might be triggered very early during their first orbital crossing, well before the galaxies merge. In what follows, we adopt the halo merger rate to model the QSO triggering rate, and we discuss the consequences of adopting the alternative subhalo merger rates in §4.2.

1.5. *BH Fueling During Mergers*

Modeling black hole fueling during mergers from first principles is not trivial: aside from the lack of physical inputs, most hydrodynamical simulations do not yet have the dynamical range to even resolve the outer Bondi radius. Although semi-analytical treatments of BH accretion are possible (e.g., Granato et al. 2004; Monaco et al. 2007), we do not consider such treatments in this paper because of the poorly understood accretion physics.

Throughout this paper we adopt the following *ansatz*:

- *QSO activity is triggered by a gas-rich major merger event ($\xi_{\rm min} < \xi \leq 1$).*
- *The build-up of the cosmic SMBH population is mainly through gas accretion during the QSO phase.*
- *The QSO light curve follows a universal form $L(L_{\rm peak},t)$, where the peak luminosity $L_{\rm peak}$ is correlated with the mass of the merged halo M_0.*

TABLE 1
NOTATION AND MODEL PARAMETERS

Symbol	Description	Value/Units
ξ	Halo mass ratio	$0 < \xi \leq 1$
M_0	Postmerger halo mass	
$M_\bullet$	Black hole mass	
$B, B_{\rm gal}$	Halo(subhalo) merger rate	$\rm Mpc^{-3} M_\odot^{-1} dz^{-1} d\xi^{-1}$
$\Phi(L,z)$	QSO luminosity function	$\rm Mpc^{-3} dL^{-1}$
B_L	QSO triggering rate	$\rm Mpc^{-3} dL^{-1} dz^{-1}$
$L_{\rm peak}$	QSO peak luminosity	
z'	QSO triggering redshift	
$f_{\rm QSO}(z, M_0)$	QSO triggering fraction	
	The $L_{\rm peak} - M_0$ relation	
γ	Slope	5/3
$C(z=0)$	Normalization at $z = 0$	$\log(6\times10^{45}) - 12\gamma$
σ_L	Scatter	0.28 dex
$C(z) = C(z=0) + \beta_1 \log(1+z)$	Evolution in normalization	$\beta_1 = 0.2\gamma = 1/3$
	The light curve model	
ϵ	Radiative efficiency	0.1
l	Eddington luminosity per $M_\odot$	$1.26\times10^{38}\,\rm ergs^{-1} M_\odot^{-1}$
$t_{\rm Salpeter}$	e-folding time	$\frac{\epsilon c^2}{l(1-\epsilon)\lambda_0}$
f	Fraction of seed BH mass to peak BH mass	10^{-3}
$t_{\rm peak}$	Time to reach the peak luminosity	$(-\ln f) t_{\rm Salpeter}$
λ_0	Eddington ratio before $t_{\rm peak}$	3
α	Power-law slope of the decaying phase	2.5
	The QSO-triggering rate	
$\xi_{\rm min}$	Minimum mass ratio	0.25
$M_{\rm min}$	Exponential lower mass cut	$3\times10^{11}\,h^{-1} M_\odot$
$M_{\rm max}(z) = M_{\rm quench}(1+z)^\beta$	Exponential upper mass cut	$10^{12}(1+z)^{3/2}\,h^{-1} M_\odot$

NOTE. — Parameter values are for the fiducial model.

Once we have specified the QSO light curve model, and estimated the QSO triggering rate from the major merger rate, we can convolve them to derive the QSO luminosity function. Within this evolutionary QSO framework, we can derive the distributions of host halo masses and BH masses for any given instantaneous luminosity and redshift, allowing us to compare with observations of quasar clustering and Eddington ratio distributions. The details of this framework are elaborated in §2.

We note that, any correlations established with our model are only for hosts where a gas-rich major merger is triggered and self-regulated BH growth occurs. Host halos which experience many minor mergers or dry mergers may deviate from these relations. We will come back to this point in the discussion section.

2. MODEL FORMALISM

In this section we describe our model setup in detail. Some notations and model parameters are summarized in Table 1 for clarity.

2.1. *Determining the Luminosity Function*

Major mergers are generally defined as events with mass ratio $\xi_{\rm min} \equiv 0.3 \lesssim \xi \leq 1$, but our framework can be easily generalized to other values of $\xi_{\rm min}$. Because it takes more than just a major merger event to trigger QSO activity, we shall model the QSO triggering rate as a redshift and mass-dependent fraction of halo merger rate. The QSO luminosity function at a given redshift z can be obtained by combining the triggering rate and the evolution of QSO luminosity[4]:

$$\Phi(L,z)dL = \int_\infty^z B_L[L_{\rm peak}(L, t_z - t_{z'}), z'] dL_{\rm peak} dz' \,, \qquad (12)$$

where $\Phi(L,z)dL$ is the QSO number density in the luminosity range $(L, L+dL)$ at redshift z; $B_L(L_{\rm peak}, z')$ is the QSO-triggering rate (merger number density per redshift per peak luminosity) at redshift z' with peak bolometric luminosity $L_{\rm peak}(L, t_z - t_{z'})$, which is determined by the specific form of the light curve; t_z and $t_{z'}$ are the cosmic time at z and z'. In integrating this equation we impose an upper limit of redshift $z_{\rm max} = 20$, but we found that the integral converges well below this redshift since the merger rate decays rapidly at high redshift. Eqn. (12) implies that the distribution of triggering redshift z' given L at redshift z is:

$$\frac{dP(L_{\rm peak}, z'|L, z)}{dz'} = p(L_{\rm peak}, z'|L, z) \propto B_L(L_{\rm peak}, z') \frac{dL_{\rm peak}}{dL} \,, \qquad (13)$$

where again $L_{\rm peak}$ is tied to L via the light curve model. Obviously only quasars with $L_{\rm peak} \geq L$ can contribute to $\Phi(L,z)$.

The QSO-triggering rate $B_L(L_{\rm peak}, z)$ is related to the halo

[4] Note that we have implicitly assumed that a second QSO-triggering merger event does not occur during the major accretion phase of the QSO – a reasonable assumption since statistically speaking bright QSOs are short-lived ($t_{\rm QSO} \ll t_H$). Observationally binary/multiple quasars within a single halo with comparable luminosities are rare occurrences ($f_{\rm binary} \lesssim 0.1\%$, e.g., Hennawi et al. 2006, 2009; Myers et al. 2008), further supporting this assumption.

merger rate $B(M_0,\xi,z)$ by:

$$B_L(L_{\rm peak},z)dL_{\rm peak} = f_{\rm QSO}(z,M_0)\int_{\xi_{\rm min}}^{1} B(M_0,\xi,z)d\xi dM_0 \ , \quad (14)$$

where we parameterize the fraction $f_{\rm QSO}$ as

$$f_{\rm QSO}(z,M_0) = \mathcal{F}(z)\exp\left[-\frac{M_{\rm min}(z)}{M_0}-\frac{M_0}{M_{\rm max}(z)}\right] , \quad (15)$$

where $\mathcal{F}(z)$ describes how $f_{\rm QSO}(z,M_0)$ decreases towards lower redshift, i.e., the overall reduction due to the consumption of cold gas. We also introduce exponential cutoffs of $f_{\rm QSO}$ at both high and low mass such that at each redshift, halos with too small M_0 cannot trigger QSO activity; on the other hand, overly massive halos cannot cool gas efficiently and BH growth halts, and the gas-rich fraction may become progressively smaller at higher halo masses (especially at low redshift). We discuss choices for these parameters later in §3.

To proceed further we must specify the relation between $L_{\rm peak}$ and M_0, and choose a light curve model. We assume that the $L_{\rm peak}-M_0$ correlation is a power-law with log-normal scatter, as motivated by some analytical arguments and hydrodynamical simulations (e.g., Wyithe & Loeb 2002, 2003; Springel et al. 2005b; Hopkins et al. 2005; Lidz et al. 2006):

$$\frac{dP(L_{\rm peak}|M_0)}{d\log L_{\rm peak}} = p(L_{\rm peak}|M_0)$$
$$= (2\pi\sigma_L^2)^{-1/2}\exp\left\{-\frac{[\log L_{\rm peak}-(C+\gamma\log M_0)]^2}{2\sigma_L^2}\right\} , \quad (16)$$

where the mean relation is:

$$\left\langle \log\left(\frac{L_{\rm peak}}{\rm erg\,s^{-1}}\right)\right\rangle = C+\gamma\log\left(\frac{M_0}{h^{-1}M_\odot}\right) , \quad (17)$$

with normalization C and power-law slope γ. The log scatter around this mean relation is denoted as σ_L. When the scatter between $L_{\rm peak}$ and M_0 is incorporated, Eqn. (14) becomes:

$$B_L(L_{\rm peak},z)$$
$$= \int\left[f_{\rm QSO}\int_{\xi_{\rm min}}^{1} B(M_0,\xi,z)d\xi\right]\frac{M_0}{L_{\rm peak}}p(L_{\rm peak}|M_0)d\log M_0 \ , \quad (18)$$

from which we obtain the probability distribution of postmerger halo mass M_0 at fixed peak luminosity $L_{\rm peak}$ and redshift z:

$$\frac{dP(M_0|L_{\rm peak},z)}{d\log M_0} = p(M_0|L_{\rm peak},z)$$
$$\propto \left[f_{\rm QSO}\int_{\xi_{\rm min}}^{1} B(M_0,\xi,z)d\xi\right]\frac{M_0}{L_{\rm peak}}p(L_{\rm peak}|M_0) \ . \quad (19)$$

We derive the probability distribution of host halo mass at given instantaneous luminosity and redshift, $p(M_0|L,z)$, as:

$$\frac{dP(M_0|L,z)}{d\log M_0} = p(M_0|L,z)$$
$$= \int\frac{dP(M_0|L_{\rm peak},z')}{d\log M_0}\frac{dP(L_{\rm peak},z'|L,z)}{dz'}dz' \ , \quad (20)$$

which can be convolved with the halo linear bias $b_M(M_0,z)$ to derive the QSO linear bias $b_L(L,z)$ at instantaneous luminosity L and redshift z.

Since we have assumed that a second QSO-triggering merger event has not occurred, at redshift z the postmerger halo M_0 should maintain most of its identity as it formed sometime earlier, i.e., we neglect mass added to the halo via minor mergers or accretion of diffuse dark matter between the halo formation time and the time when the QSO is observed at redshift z.

Given the halo mass distribution (20), we can further derive the "active" halo mass function hosting a QSO with $L > L_{\rm min}$:

$$\frac{d\Psi_{M_0}}{d\log M_0} = \int_{L_{\rm min}}^{\infty}\Phi(L,z)dL\frac{dP(M_0|L,z)}{d\log M_0} \ , \quad (21)$$

which can be used to compute the halo duty cycles.

2.2. *The Light Curve Model*

For the light curve model there are several common choices in merger-based cosmological QSO models (e.g., Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003): a) a light bulb model in which the QSO shines at a constant value $L_{\rm peak}$ for a fixed time $t_{\rm QSO}$; b) an exponential decay model $L = L_{\rm peak}\exp(-t/t_{\rm QSO})$; c) a power-law decay model $L = L_{\rm peak}(1+t/t_{\rm QSO})^{-\alpha}$ with $\alpha > 0$. Once a light curve model is chosen, the total accreted mass during the QSO phase is simply[5]:

$$M_{\bullet,\rm relic} = \frac{1-\epsilon}{\epsilon c^2}\int_0^{\infty} L(L_{\rm peak},t)dt \ . \quad (22)$$

The evolution of the *luminosity* Eddington ratio $\lambda \equiv L/L_{\rm Edd}$ is (neglecting the seed BH mass):

$$\lambda(t) = \frac{L(L_{\rm peak},t)}{lM_\bullet(t)} = \frac{L(L_{\rm peak},t)}{\frac{l(1-\epsilon)}{\epsilon c^2}\int_0^t L(L_{\rm peak},t')dt'} , \quad (23)$$

where $l \equiv 1.26\times10^{38}\,{\rm erg\,s^{-1}}M_\odot^{-1}$ is the Eddington luminosity per $M_\odot$.

Unfortunately none of the three LC models is self-consistent without having $\lambda \gg 1$ at the very beginning of BH growth, simply because they neglect the rising part of the light curve. To remedy this, we must model the evolution of luminosity and BH growth self-consistently (e.g., Small & Blandford 1992; Yu & Lu 2004, 2008). We consider a general form of light curve where the BH first grows exponentially at constant *luminosity* Eddington ratio λ_0 (e.g., Salpeter 1964) to $L_{\rm peak}$ at $t = t_{\rm peak}$, and then the luminosity decays monotonically as a power-law (e.g., Yu & Lu 2008):

$$L(L_{\rm peak},t) = \begin{cases} L_{\rm peak}\exp\left[\frac{l(1-\epsilon)\lambda_0}{\epsilon c^2}(t-t_{\rm peak})\right], & 0\le t\le t_{\rm peak} \\ L_{\rm peak}\left(\frac{t}{t_{\rm peak}}\right)^{-\alpha}, & t\ge t_{\rm peak} \ , \end{cases} \quad (24)$$

where $t_{\rm peak}$ is determined by:

$$L_{\rm peak} = l\lambda_0 M_{\bullet,0}\exp\left[\frac{l(1-\epsilon)\lambda_0}{\epsilon c^2}t_{\rm peak}\right] , \quad (25)$$

[5] Throughout the paper we assume constant radiative efficiency ϵ. At the very late stage of evolution or under certain circumstance (i.e., hot gas accretion within a massive halo), a SMBH may accrete via *radiatively-inefficient accretion flows* (RIAFs) with very low ϵ and mass accretion rate (e.g., Narayan & Yi 1995). We neglect this possible accretion state since the mass accreted during this state is most likely negligible (e.g. Hopkins et al. 2006).

where $M_{\bullet,0}$ is the seed BH mass at the triggering time $t = 0$. In eqn. (24), α determines how rapidly the LC decays. Larger α values lead to more rapid decay. Thus this model accommodates a broad range of possible light curves. With this light curve model (24) we have:

$$M_\bullet(L_{\rm peak},t) = \begin{cases} \dfrac{L_{\rm peak}}{l\lambda_0}\exp\left[\dfrac{l(1-\epsilon)\lambda_0}{\epsilon c^2}(t-t_{\rm peak})\right], & 0\le t\le t_{\rm peak} \\ \dfrac{L_{\rm peak}}{l\lambda_0}+\dfrac{(1-\epsilon)L_{\rm peak}t_{\rm peak}}{\epsilon c^2}\times & \\ \dfrac{1}{1-\alpha}\left[\left(\dfrac{t}{t_{\rm peak}}\right)^{1-\alpha}-1\right], & t\ge t_{\rm peak}\ , \end{cases} \quad (26)$$

where we require $\alpha > 1$ so that a BH cannot grow infinite mass. The evolution of the luminosity Eddington ratio is therefore:

$$\lambda(t) = \begin{cases} \lambda_0, & 0\le t\le t_{\rm peak} \\ \dfrac{(t/t_{\rm peak})^{-\alpha}}{\dfrac{1}{\lambda_0}+\dfrac{l(1-\epsilon)t_{\rm peak}}{\epsilon c^2(1-\alpha)}\left[\left(\dfrac{t}{t_{\rm peak}}\right)^{1-\alpha}-1\right]}, & t\ge t_{\rm peak}\ , \end{cases} \quad (27)$$

where during the decaying part of the LC the Eddington ratio monotonically decreases to zero. We assume that the seed BH mass is a fraction f of the peak mass $M_\bullet(t_{\rm peak}) \equiv L_{\rm peak}/(l\lambda_0)$, and we have

$$t_{\rm peak} = -\frac{\epsilon c^2}{l(1-\epsilon)\lambda_0}\ln f = (-\ln f)t_{\rm Salpeter}\ . \quad (28)$$

In what follows we set $f = 10^{-3}$. Thus the seed BH is negligible compared to the total mass accreted, and $t_{\rm peak} \approx 6.9 t_{\rm Salpeter}$ where $t_{\rm Salpeter}$ is the (e-folding) Salpeter timescale (Salpeter 1964). Although $\lambda_0 = 1$ is the formal definition of Eddington-limited accretion, it assumes the electron scattering cross section, and in practice super-Eddington accretion cannot be ruled out (e.g., Begelman 2002). So we consider λ_0 within the range $\log\lambda_0 \in [-1, 1]$.

The relic BH mass, given eqn. (26), is simply

$$M_{\bullet,\rm relic} = \frac{L_{\rm peak}}{l\lambda_0}+\frac{(1-\epsilon)L_{\rm peak}t_{\rm peak}}{(\alpha-1)\epsilon c^2} = \frac{L_{\rm peak}}{l\lambda_0}\left[1-\frac{\ln f}{\alpha-1}\right]\ . \quad (29)$$

If instead we choose an exponential model for the decaying part of the LC (e.g., Kauffmann & Haehnelt 2000), $L(t > t_{\rm peak}) = L_{\rm peak}\exp[-(t-t_{\rm peak})/t_{\rm QSO}]$, then the relic mass becomes:

$$M_{\bullet,\rm relic} = \frac{L_{\rm peak}}{l\lambda_0}+\frac{(1-\epsilon)L_{\rm peak}t_{\rm QSO}}{\epsilon c^2}\ . \quad (30)$$

Therefore LC models with large values of α mimic the exponentially decaying model, and we do not consider the exponential LC model further.

If the decaying parameter α is independent of mass, then Eqns. (26) and (29) imply $M_{\bullet,\rm relic} \propto L_{\rm peak}$. In other words, the scalings between $L_{\rm peak}$ and M_0, and between $M_{\bullet,\rm relic}$ and M_0 should be the same for QSO-triggering merger remnants (e.g., early type galaxies). Some theoretical models and simulations predict $L_{\rm peak} \propto M_0^{4/3}$ (e.g., King 2003; Di Matteo et al. 2005; Springel et al. 2005a), where momentum is conserved in self-regulated feedback; while others invoking energy conservation predict $L_{\rm peak} \propto M_0^{5/3}$ (e.g., Silk & Rees 1998; Wyithe & Loeb 2003). Two recent determinations of the local black hole mass-halo mass relation gave $M_{\bullet,\rm relic} \propto M_0^{1.65}$ (Ferrarese 2002) and $M_{\bullet,\rm relic} \propto M_0^{1.27}$ (Baes et al. 2003). Within uncertainties $M_{\bullet,\rm relic} \propto L_{\rm peak}$ seems to be a reasonable scaling, therefore we do not consider further mass dependence of α. Given the definition of $t_{\rm peak}$ (Eqn. 28), our light curve model (24) is then *universal* in the sense that it scales with $L_{\rm peak}$ in a self-similar fashion.

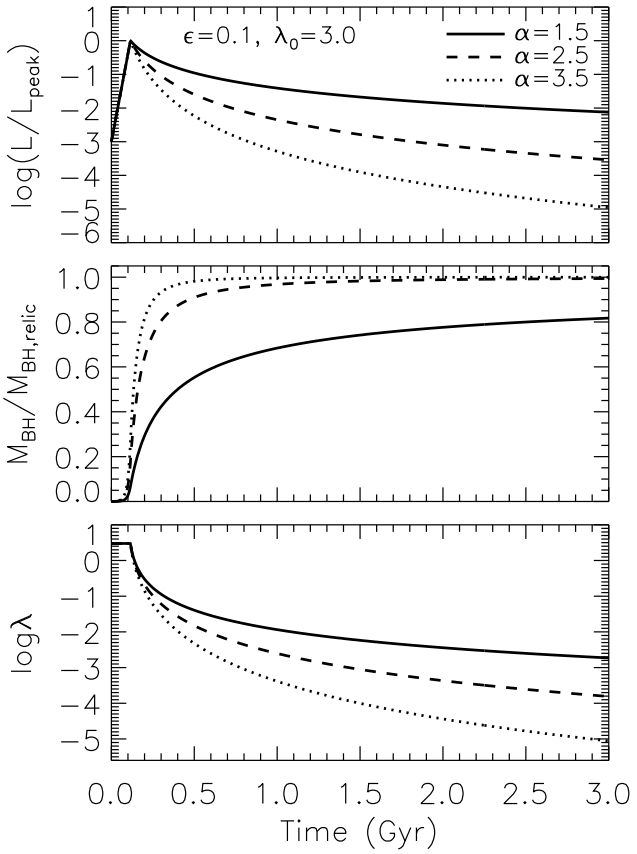

FIG. 2.— Example light curves for $\epsilon = 0.1$ and $\lambda_0 = 3.0$, for $\alpha = 1.5, 2.5, 3.5$. Larger values of α lead to more rapid decay. The time that the QSO spends above half of its peak luminosity is short, < 100 Myr for large values of α.

As an example, Fig. 2 shows three light curves with $\alpha = 1.5, 2.5, 3.5$ for $\epsilon = 0.1$ and $\lambda_0 = 3.0$, in which case $t_{\rm peak} \sim 115$ Myr. The time that a QSO spends above 50% of its peak luminosity is typically $\lesssim 100$ Myr for sharp decaying curves, but it can be substantially longer for extended decaying curves.

2.3. *BH Mass and Eddington Ratio Distributions*

Now we have specified the light curve model and the peak luminosity-halo mass relation, and tied the luminosity function $\Phi(L, z)$ to the QSO-triggering merger rate. We continue to derive the instantaneous BH mass and Eddington ratio distributions at fixed instantaneous luminosity L and redshift z, which can be compared with observationally determined distributions (e.g., Babić et al. 2007; Kollmeier et al. 2006; Shen et al. 2008a; Gavignaud et al. 2008).

Suppose that at redshift z we observe quasars with instantaneous luminosity L. These quasars consist of objects triggered at different earlier redshift z' and are at different stages of their evolution when witnessed at z (e.g., Eqn. 12). There is a characteristic earlier redshift z_c, determined by

$$t_{\rm age}(z) - t_{\rm age}(z_c) = t_{\rm peak}\ , \quad (31)$$

where $t_{\rm age}$ is the cosmic time. Quasars triggered between $[z_c, z]$ are all in the rising part of the LC; while quasars trig-

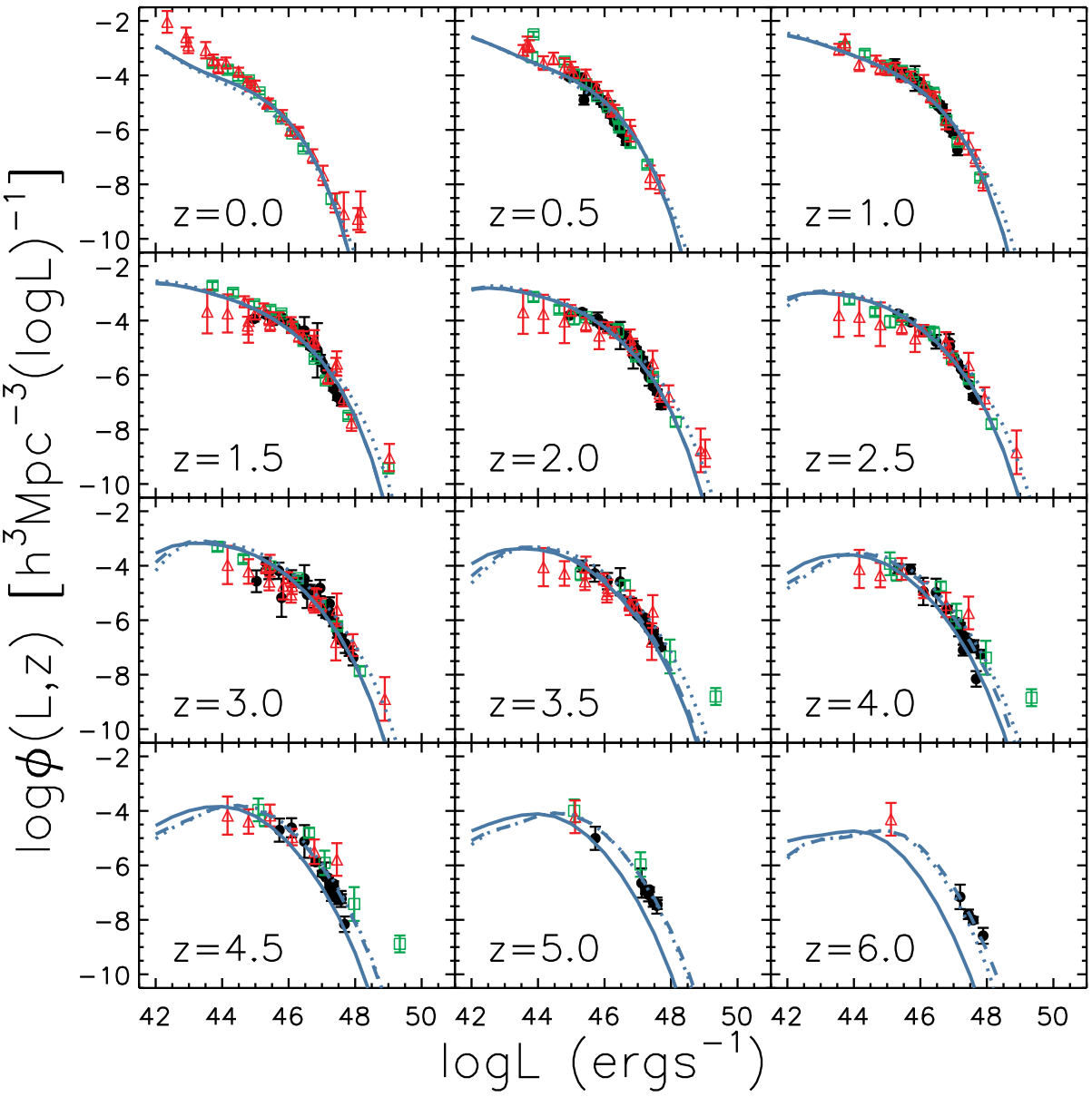

FIG. 3.— Predicted bolometric luminosity functions $\phi(L,z) \equiv d\Psi/d\log L$ in our fiducial model (solid lines). Overplotted are the complied bolometric LF data from Hopkins et al. (2007) for optical (black circles), soft X-ray (green squares) and hard X-ray (red triangles) samples. The dashed lines are the predicted LF for a model where the normalization in the mean $L_{\rm peak}-M_0$ relation is reduced by an additional amount at $z>3.5$, tuned to fit the LF at $z\gtrsim 4.5$ (see §4.2 for details). The blue dotted lines are the predictions based on an alternative prescription for the redshift evolution in the normalization of the $L_{\rm peak}-M_0$ relation as discussed in §4.4.

gered earlier than z_c are all in the decaying part of the LC. In order to contribute to $\Phi(L,z)$, a quasar should have peak luminosity $L_{\rm peak}=L$ if it is triggered right at z_c, and it should have $L_{\rm peak}>L$ otherwise. Therefore the triggering redshift distribution $dP(L_{\rm peak},z'|L,z)/dz'$ (Eqn. 13) peaks around[6] $z'\approx z_c$ since $B_L(L_{\rm peak},z')$ decreases rapidly when $L_{\rm peak}$ increases. Moreover, at the observing redshift z, all the contributing quasars triggered between $[z_c,z]$ will have Eddington ratio $\lambda_{L,z'\rightarrow z}=\lambda_0$ and BH mass $M_{\bullet,L,z'\rightarrow z}=L/(l\lambda_0)$; while all the contributing quasars triggered before z_c will have Eddington ratio $\lambda_{L,z'\rightarrow z}<\lambda_0$ and BH mass $M_{\bullet,L,z'\rightarrow z}>L/(l\lambda_0)$. The probability distribution of instantaneous BH mass $M_\bullet$ at the observing redshift z and luminosity L is therefore:

$$\frac{dP(M_\bullet|L,z)}{d\log M_\bullet}=\frac{\frac{dP}{dz'}\frac{dz'}{d\log M_\bullet}}{1-\int_{z_c}^{z}\frac{dP}{dz'}dz'}\,,\qquad M_\bullet\geq\frac{L}{l\lambda_0}\,, \tag{32}$$

where $dP(z'|L,z)/dz'$ is given by Eqn. (13) and $dz'/dM_\bullet$ is determined by *the decaying half* of the LC (Eqn. 27), i.e., the triggering redshift $z'(\geq z_c)$ is a monotonically increasing function of $M_\bullet$ (note that here $M_\bullet$ refers to the instantaneous BH mass observed at z). The denominator in the above equation is to normalize the distribution – we literally augment the distribution with the pileup of objects with $M_\bullet=L/(l\lambda_0)$ triggered within $[z_c,z]$. The fraction of such objects is about 20% for $L\gtrsim 10^{47}\,{\rm ergs^{-1}}$ and becomes negligible at lower luminosities. This procedure removes the spike (δ function) at the constant

[6] Except for low luminosities when the lower halo mass cut $M_{\rm min}$ in the QSO-triggering rate starts to kick in, the distribution $dP(L_{\rm peak},z'|L,z)/dz'$ instead has a dip around z_c.

BH mass $M_\bullet = L/(l\lambda_0)$ in the distribution, which is an artifact of our model[7]. Eqn. (32) gives the equivalent distribution of Eddington ratios at z and at instantaneous luminosity L.

Given the luminosity function (12) and the distribution of $M_\bullet$ (32) we can determine the *active* BH mass function (i.e., the mass function of active QSOs above some luminosity cut $L_{\rm min}$) as

$$\frac{d\Psi_{M_\bullet}}{d\log M_\bullet} = \int_{L_{\rm min}}^{\infty} \Phi(L,z) dL \frac{dP(M_\bullet|L,z)}{d\log M_\bullet} . \quad (33)$$

3. THE REFERENCE MODEL

We are now ready to convolve the LC model with the QSO-triggering rate Eqn. (18) to predict the LF $\Phi(L,z)$ using Eqn. (12). Before we continue, let us review the model parameters and available observational/theoretical constraints.

- *QSO triggering rate [$\xi_{\rm min}$, $M_{\rm min}(z)$, $M_{\rm max}(z)$, $\mathcal{F}(z)$]:* We choose our fiducial value of $\xi_{\rm min} \approx 0.3$ following the traditional definition of major mergers (1 : 3 or 1 : 4). For the minimum halo mass $M_{\rm min}$ below which efficient accretion onto the BH is hampered we simply choose $M_{\rm min}(z) = 3\times10^{11}\,h^{-1}M_\odot$. We model the maximum halo mass cut above which the gas-rich merger fraction drops as a function of redshift:

$$M_{\rm max}(z) = M_{\rm quench}(1+z)^\beta , \quad (34)$$

with $\beta > 0$. Therefore at lower redshift halos have a smaller upper threshold. Since at $z < 0.5$ rich group to cluster size halos no longer host luminous QSOs we tentatively set $M_{\rm quench} = 1\times10^{12}\,h^{-1}M_\odot$. The parameters $M_{\rm min}$ and $M_{\rm max}$ control the the shape of the LF at both the faint and bright luminosity ends. The global gas-rich merger fraction, $\mathcal{F}(z)$, is more challenging to determine. Direct observations of the cold gas fraction as function of redshift and stellar mass (and therefore halo mass) are still limited by large uncertainties. Moreover, how *gas-rich* a merger needs to be in order to trigger efficient BH accretion is not well constrained. For these reasons, we simply model $\mathcal{F}(z)$ as a two-piece function such that $\mathcal{F}(z) = 1$ when $z > 2$, the peak of the bright quasar population; and $\mathcal{F}(z)$ linearly decreases to $\mathcal{F}_0$ at $z = 0$. We note again that $f_{\rm QSO}(z, M_0)$ should be regarded as the fraction of major mergers that trigger QSO activity.

- *The $L_{\rm peak}-M_0$ relation (C, γ, σ_L)*: some merger event simulations (e.g., Springel et al. 2005a; Hopkins et al. 2005; Lidz et al. 2006) reveal a correlation between the peak luminosity and the mass of the (postmerger) halo: $\langle L_{\rm peak}\rangle \approx 3\times10^{45}(M_0/10^{12}\,h^{-1}M_\odot)^{4/3}\,{\rm ergs^{-1}}$ at $z = 2$, with a lognormal scatter $\sigma_L = 0.35\,$dex. Other analytical arguments predict $\gamma = 5/3$, where energy is conserved during BH feedback (e.g., Silk & Rees 1998; Wyithe & Loeb 2003). We choose γ values between $[4/3, 5/3]$ since the above simulations are also consistent with the $\gamma = 5/3$ slope. Furthermore, we allow the normalization parameter C to evolve with redshift, $C(z) = C(z=0) + \log[(1+z)^{\beta_1}]$ with $\beta_1 > 0$. Thus for the same peak luminosity, the host halos become less massive at higher redshift. This decrease of the characteristic halo mass with redshift is also modeled by several other authors, although we do not restrict to $\beta_1 = 1$ (e.g., Lapi et al. 2006), or more complicated prescriptions (e.g., Wyithe & Loeb 2003; Croton 2009), since the form of the $L-M_0$ (or $M_\bullet-M_0$) relation is also different in various prescriptions. A smaller characteristic halo mass is easier to account for the QSO abundance at high redshift since there are more mergers of smaller halos, but it also reduces the clustering strength. The intrinsic scatter of the $L_{\rm peak}-M_0$ relation, σ_L, will have effects on both the luminosity function and clustering. Larger values of σ_L lead to higher QSO counts, but will also dilute the clustering strength due to the up-scattering of lower mass halos (e.g., White et al. 2008). The $L_{\rm peak}-M_0$ relation establishes a baseline for the mapping from halos to SMBHs. Although in our fiducial model we adopt the above rather simple scaling [$\propto(1+z)^{\beta_1}$] for the redshift evolution in $C(z)$, we will discuss alternative parameterizations for $C(z)$ in §4.2 and §4.4.

- *The light curve model (λ_0, ϵ, $t_{\rm peak}$, α):* our chosen fiducial value for $t_{\rm peak}$ is $t_{\rm peak} = 6.9t_{\rm Salpeter}$ ($f = 10^{-3}$); but our results are insensitive to the exact value of f. The radiative efficiency is $\epsilon \approx 0.1$ from the Soltan argument (Eqn. 1). We choose λ_0 between $[0.1, 10]$ and $\alpha > 1.1$.

Normally to find the best model parameters one needs to perform a χ^2 minimization between model predictions and observations. We do not perform such exercise here because it is difficult to assign relative weights to different sets of observations (i.e., LF, quasar clustering, Eddington ratio distributions, etc), and we don't know well enough the systematics involved. Instead, we experiment with varying the model parameters within reasonable ranges, to achieve a global "good" (as judged by eye) fit to the overall observations. Our fiducial model has the following parameter values: $\xi_{\rm min} = 0.25$, $\mathcal{F}_0 = 1.0$, $\beta = 1.5$, $\gamma = 5/3$, $C(z=0) = \log(6\times10^{45}) - 12\gamma$, $\beta_1 = 0.2\gamma$, $\sigma_L = 0.28$, $\lambda_0 = 3.0$ and $\alpha = 2.5$, which are all within reasonable ranges. In particular we found that the parameterization of $\mathcal{F}(z)$ is unnecessary because the effect of cold gas consumption is somewhat described by $M_{\rm max}(z)$ already. Below we describe in detail the predictions of this reference model and comparison with observations, and we defer the model variants and caveats to §4.

Fig. 3 shows the model LF $d\Psi/d\log L \equiv L\ln(10)\Phi(L,z)$ in solid lines, given by Eqn. (12), where we overplot the compiled bolometric luminosity function data in Hopkins et al. (2007). In computing Eqn. (12) we integrate up to $z = 20$ but the integral converges well before that. The model under-predicts the counts at the faint luminosity end ($L < 10^{45}\,{\rm ergs^{-1}}$) for $z < 0.5$, leaving room for low luminosity AGNs triggered by mechanisms other than a major merger event (i.e., by secular processes). At redshift $z \gtrsim 4.5$, the model also underestimates the QSO abundance. This could be alleviated if the characteristic halo mass shifts to even lower values, i.e., even larger values of the parameter C at $z \gtrsim 4.5$ (see discussions in §4.2 and §4.4). Alternatively, the high-z SMBH population may be different in the sense that it is not tied to $\xi \gtrsim 0.3$ major mergers events directly. Also, the halo merger rate (Eqn. 4) has been extrapolated to such high redshifts for the massive halos considered here. Despite these

[7] We have tested with the alternative BH mass distribution at instantaneous luminosity L and redshift z where the contribution of objects with $M_\bullet = L/(l\lambda_0)$ is described by a δ function with normalization $\int_{z_c}^{z}\frac{dP}{dz'}dz'$, and we find that other derived distributions based on this BH mass distribution are almost identical to those using Eqn. (32).

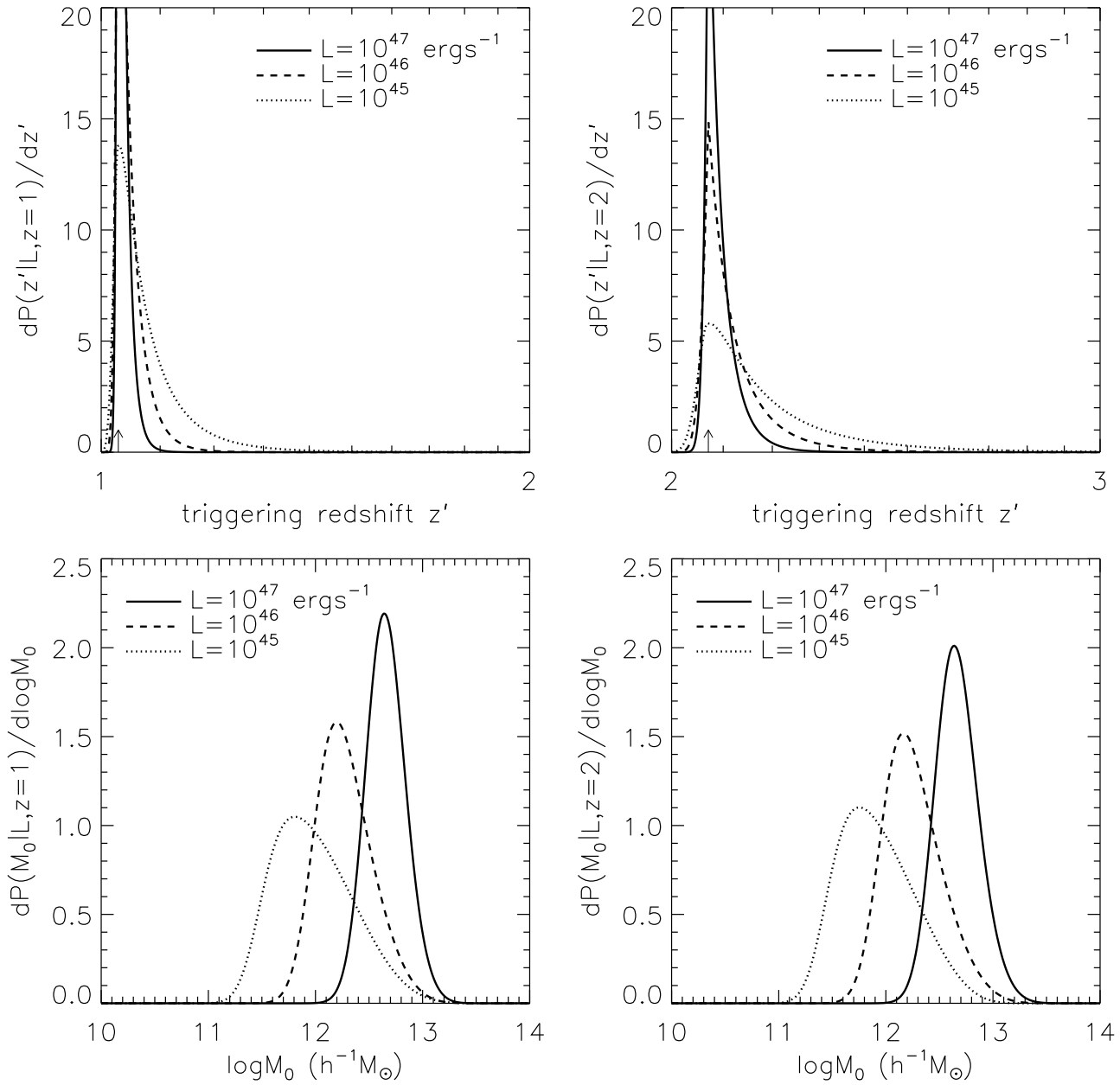

FIG. 4.— Distributions of triggering redshift (upper panels) and halo mass (lower panels) in our fiducial model, for three instantaneous luminosities and two values of observing redshifts. The distributions of triggering redshift generally peak around z_c (marked by arrows) given by Eqn. (31). For lower luminosity, the distribution of the triggering redshift is more extended. As a consequence, there is a significant population of faded low-luminosity QSOs which have massive hosts.

facts, the model correctly reproduces the LF from $z \approx 0.5$ to $z \approx 4.5$. The turnover below $L \approx 10^{44}\,\mathrm{ergs}^{-1}$ at $z \gtrsim 2.5$ is caused by the lower mass cutoff $M_{\rm min} = 3\times10^{11}\,h^{-1}M_\odot$. At lower redshift, this turnover flattens due to the gradual pile up of evolved high-peak luminosity quasars well after $t_{\rm peak}$.

We show some of the predicted distributions in Fig. 4 for this reference model. The upper panels show the distributions of the triggering redshift z' for instantaneous luminosities $L = 10^{45}, 10^{46}, 10^{47}\,\mathrm{ergs}^{-1}$ at $z = 1$ (upper left) and $z = 2$ (upper right). As expected, the distribution of the triggering redshift peaks around the characteristic redshift z_c given by Eqn. (31). The bottom panels show the distributions of host halo mass M_0 for the three instantaneous luminosities, at the two redshifts respectively. For bright quasars ($L > 10^{46}\,\mathrm{ergs}^{-1}$), the distributions of M_0 are roughly log-normal, with the width and mean determined mainly through the $L_{\rm peak} - M_0$ relation – but also slightly modified through the convolutions with other distributions (see Eqns. 13, 19 and 20). For faint QSOs ($L \lesssim 10^{45}\,\mathrm{ergs}^{-1}$), however, the halo mass distribution has a broad high-mass tail contributed by evolved high-peak luminosity quasars triggered earlier. This contamination of massive halos hosting low luminosity QSOs will increase the clustering bias of the faint quasar population.

Fig. 5 shows our model predictions for the redshift and luminosity dependence of quasar bias, compared with observations. The model-predicted redshift evolution of quasar bias is in good agreement with observations (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Myers et al. 2006, 2007a,b; Shen et al. 2007; da Ângela et al. 2008; Padmanabhan et al.

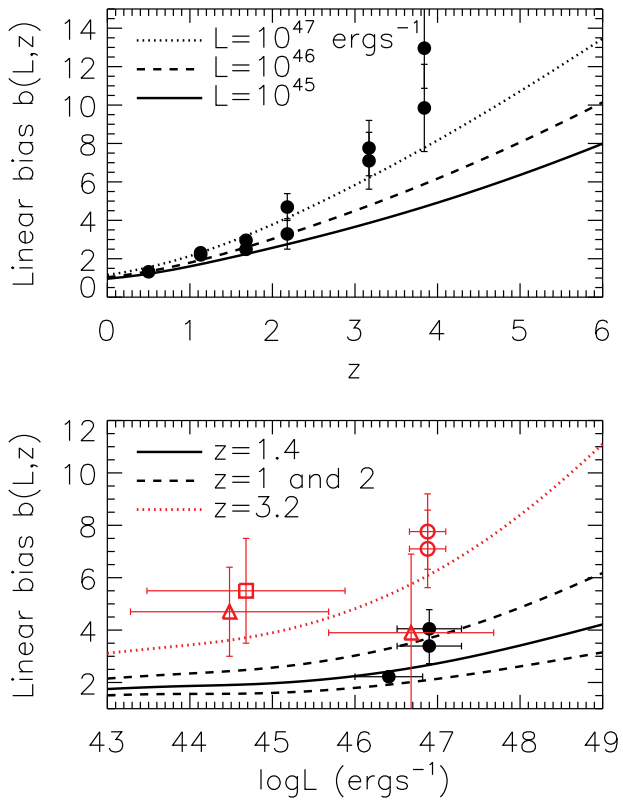

FIG. 5.— Predictions for the linear bias of quasar clustering in our fiducial model. *Upper*: bias evolution with redshift. The data points are from the measurements in Shen et al. (2009) and three lines show the predicted bias for quasars with instantaneous luminosity $L = 10^{45}, 10^{46}, 10^{47}$ ergs^{-1}. The median luminosity of quasars used in the clustering analysis for the six redshift bins are: $\langle \log(L/\mathrm{ergs}^{-1}) \rangle = 45.6, 46.3, 46.6, 46.9, 46.9, 47.1$. *Bottom:* predicted luminosity dependence of quasar bias. The filled circles are from the measurements in Shen et al. (2009) for the 10% most luminous and the remainder of a sample of quasars spanning $0.4 < z < 2.5$. The open circles are the measurement of quasar clustering for $2.9 < z < 3.5$ in Shen et al. (2009). The open square is from Francke et al. (2008) using cross-correlation of X-ray selected AGN with high redshift galaxies, and the open triangles are from Adelberger & Steidel (2005) using cross-correlation of optical AGNs with galaxies. Note that for the quasar clustering data we plot biases derived from both with and without including negative correlation function data in the fitting (see table 1 of Shen et al. 2009).

2008; Shen et al. 2008b, 2009; Ross et al. 2009). But it has some difficulties in accounting for the large bias at $z \approx 4$, resulting from the need to reproduce the quasar abundance at such high redshift (i.e., the characteristic halo mass shifts to lower values). We discuss possible solutions to this problem in §4.2. Our model also predicts the luminosity dependence of quasar clustering. At low luminosities, quasar bias depends weakly on luminosity, which is consistent with observations (e.g., Porciani & Norberg 2006; Myers et al. 2007a; da Ângela et al. 2008; Shen et al. 2009). This simply reflects the fact that quasars are not light bulbs, i.e., there is non-negligible scatter around the mean $L_{\rm peak} - M_0$ relation, and some faint quasars live in massive halos because of the evolving light curve (e.g., Adelberger & Steidel 2005; Hopkins et al. 2005; Lidz et al. 2006). On the other hand, the model predicts that the bias increases rapidly towards high luminosities and/or at high redshift. This is consistent with the findings by Shen et al. (2009) that the most luminous quasars cluster more strongly than intermediate luminosity quasars at $z < 3$. Our model-predicted luminosity dependent quasar clustering broadly agrees with the measurements in Shen et al. (2009) for $0.4 < z < 2.5$, but we caution that their measurements are done for samples with a broad redshift and luminosity range in order to build up statistics (e.g., see their fig. 2), hence a direct comparison is somewhat difficult. At redshift ~ 3, Adelberger & Steidel (2005) and Francke et al. (2008) measured the clustering of low luminosity AGN using cross-correlation with galaxies. Their data are shown in Fig. 5 in open square and triangles for the measurements in Francke et al. (2008) and Adelberger & Steidel (2005) respectively, which are broadly consistent with our predictions. However the clustering of optically bright quasars is apparently at odds with the low bias value derived from the AGN-galaxy cross-correlation in Adelberger & Steidel (2005), likely caused by the fact that the latter AGN sample spans a wide redshift range $1.6 \lesssim z \lesssim 3.7$ and luminosity range, and that the uncertainty in their bias determination is large.

Fig. 6 shows our model predictions for the Eddington ratio distributions at various redshifts and instantaneous luminosities. Aside from the cutoff at $\lambda_0 = 3$ (our model setup), these Eddington ratio distributions are approximately log-normal, and broaden towards fainter luminosities. This again reflects the nature of the light curve – at lower luminosities, there are more objects at their late evolutionary stages and shining at lower Eddington ratios. These predictions are in good agreement with the observed Eddington ratio distributions for bright quasars ($L \gtrsim 10^{46}$ ergs^{-1}) where the distribution is narrow and peaks at high mean values ($\langle \log\lambda \rangle \in [-1, 0]$) (e.g., Woo & Urry 2002; Kollmeier et al. 2006; Shen et al. 2008a), as well as for faint AGNs ($L \lesssim 10^{45}$ ergs^{-1}) where the distribution is broader and peaks at lower mean values ($\langle \log\lambda \rangle \in [-3, -1]$) (e.g., Babić et al. 2007; Gavignaud et al. 2008). However the predicted mean values and widths are not necessarily exactly the same as those determined from observations, since uncertainties in the BH mass estimators used in these observations may introduce additional scatter and biases in the observed Eddington ratio distributions (e.g., Shen et al. 2008a). The Eddington ratio distributions of faint AGNs ($L \lesssim 10^{44}$ ergs^{-1}) are particularly broad at low redshift since more high-$L_{\rm peak}$ objects have had enough time to evolve. On the other hand, at high redshift $z \gtrsim 2$, the minimum allowable Eddington ratio is set by the cosmic age at that redshift, e.g., $\lambda_{\rm min} \approx 10^{-4}$ at $z = 2$ using Eqn. (27). This explains the narrowing of the lower end of the Eddington ratio distribution for $L = 10^{44}$ ergs^{-1}, seen in the right panel of Fig. 6.

Since we have reproduced the observed luminosity function and predicted the Eddington ratio distributions as function of luminosity, we can use Eqn. (33) to derive the *active* BHMF in QSOs above some minimum luminosity. In Fig. 7 we show the *total* BHMF (setting $L_{\rm min} = 0$ in Eqn. 33) assembled at several redshifts, where the gray shaded region indicates the estimates of the local dormant BHMF based on various galaxy bulge-BH scaling relations (Shankar et al. 2009b). Our model prediction for the local BHMF is incomplete at the low-mass end, mainly because our model does not include low luminosity AGN activity (presumably linked to $M_\bullet \lesssim 10^7 M_\odot$ BH growth) possibly triggered by secular processes (see further discussion in §4.2), and also because we set the minimal halo mass that can trigger QSO activity during major mergers $M_{\rm min} = 3 \times 10^{11} h^{-1} M_\odot$ in order to reproduce the faint-end LF at high redshift. At the high mass end, the predicted slope in the *total* BHMF broadly agrees with (although is shallower than) that for the local dormant BHMF. The majority of the present-day $\gtrsim 10^{8.5} M_\odot$ BHs were already in place by $z = 1$,

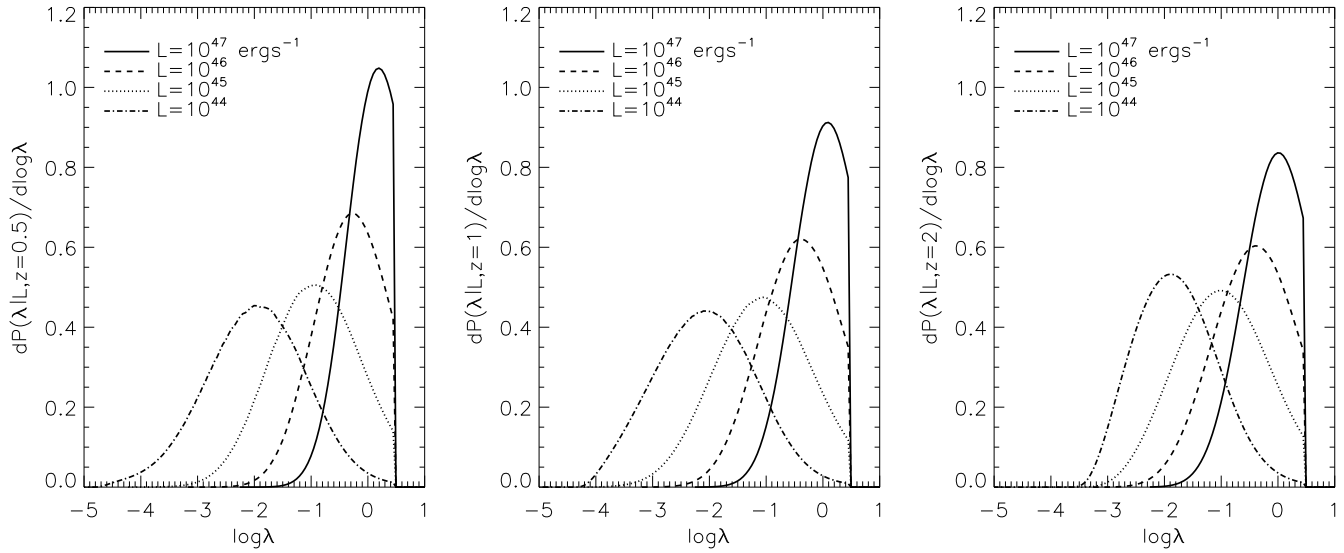

FIG. 6.— Predictions for the Eddington ratio distributions in our fiducial model for three different redshifts. The distribution is narrower and has higher peak values towards higher luminosities. These distributions are in good agreement with the observed distributions for both bright quasars and low luminosity AGNs (e.g., Kollmeier et al. 2006; Babić et al. 2007; Shen et al. 2008a; Gavignaud et al. 2008).

but less than 50% of them were assembled by $z = 2$. This is somewhat in disagreement with McLure & Dunlop (2004), who claimed that the majority of $> 10^{8.5}\,M_\odot$ SMBHs are already in place at $z \sim 2$ based on virial BH mass estimates of optically selected bright quasars. We suspect this discrepancy is caused by: 1) the fact that virial BH mass estimates tend to systematically overestimate the true BH masses due to a Malmquist-type bias (see discussions in Shen et al. 2008a); 2) the high-mass end slope in the local dormant BHMF is steeper than our model predictions. On the other hand, our model predictions are in good agreement with the predictions by Shankar et al. (2009b) for the high-mass end. It is possible to make our model prediction agree with the local BHMF at the high-mass end by imposing some cutoff in the light curve [Eqn. (24)] such that BHs cannot grow too massive; but a more accurate observational determination of the high-mass end of the local BHMF is needed to resolve these issues.

Fig. 8 shows the halo duty cycles, defined as the ratio of the number density of active halos hosting QSOs brighter than $L_{\rm min}$ to that of all halos, as function of halo mass. We have used the Sheth et al. (2001) halo mass function and Eqn. (21) for the active halo mass function. For quasars ($L > 10^{45}\,{\rm ergs^{-1}}$) and for typical halo mass $M_0 \sim 2\times10^{12}\,h^{-1}M_\odot$, the duty cycle is $\sim 0.15, 0.1, 0.03, 0.01$ at $z = 3, 2, 1, 0.5$.

We can also compare the model predicted BH mass distributions for quasars within certain luminosity ranges with the virial BH mass estimates. For this purpose we use the virial BH mass estimates from Shen et al. (2008a) for the SDSS DR5 quasar catalog (Schneider et al. 2007). These optical quasars are flux-limited to $i = 19.1$ at $z \lesssim 3$ and we have used eqn. (1) in Shen et al. (2009) to convert i-band magnitude to bolometric luminosity. For quasars/AGNs where the SMBHs are still actively accreting, what we observe is the instantaneous BH mass rather than the relic BH mass. In Fig. 9 the solid lines show the model predictions for the BH mass distributions of SDSS quasars at $z = 1$ and $z = 2$, weighted by the LF; the dashed lines show the distributions based on virial BH masses. It is remarkable that not only the distributions of our model predictions are broader, but also the peaks are shifted to lower masses (~ 0.6 dex) compared with those from virial mass estimates – as already discussed extensively in Shen et al. (2008a, i.e., the Malmquist-type bias in virial mass estimates).

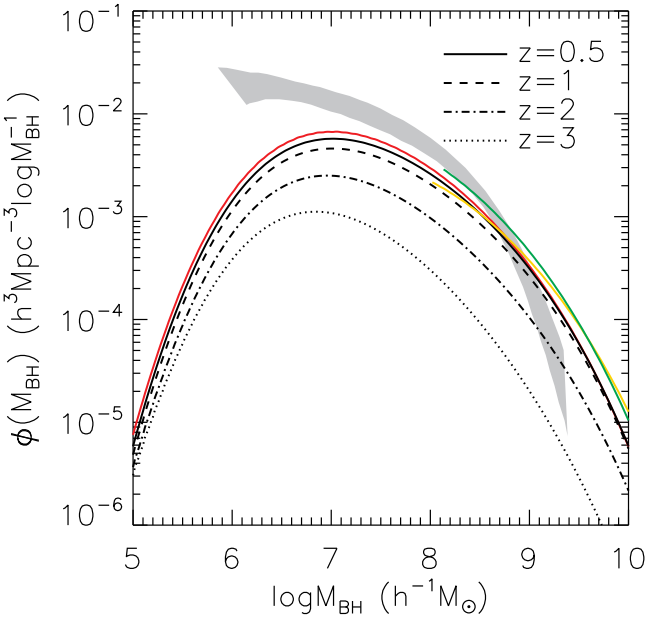

FIG. 7.— BHMFs assembled at various redshifts in our fiducial model. The gray shaded region shows the estimates for the local BHMF from Shankar et al. (2009b). The red line shows the prediction for the local BHMF, which is incomplete at $M_\bullet \lesssim 10^{7.5}\,M_\odot$ by a factor of a few because we did not include contributions from AGNs triggered by secular processes or minor mergers (as reflected in the failure to reproduce the LF at the low luminosity end $L < 10^{45}\,{\rm ergs^{-1}}$ at $z < 0.5$, see Fig. 3). The yellow and green lines show the predicted local BHMF at $M_\bullet > 10^8\,h^{-1}M_\odot$ after correcting for BH coalescence; these corrections are likely upper limits (see §4.3 for details).

4. DISCUSSION

4.1. *Comparison with Previous Work*

The merger basis of our framework, as advocated by many authors (e.g., Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003; Volonteri et al. 2003), provides a physical origin for the QSO population, and distinguishes the current study from other works which focus on BH growth using the QSO

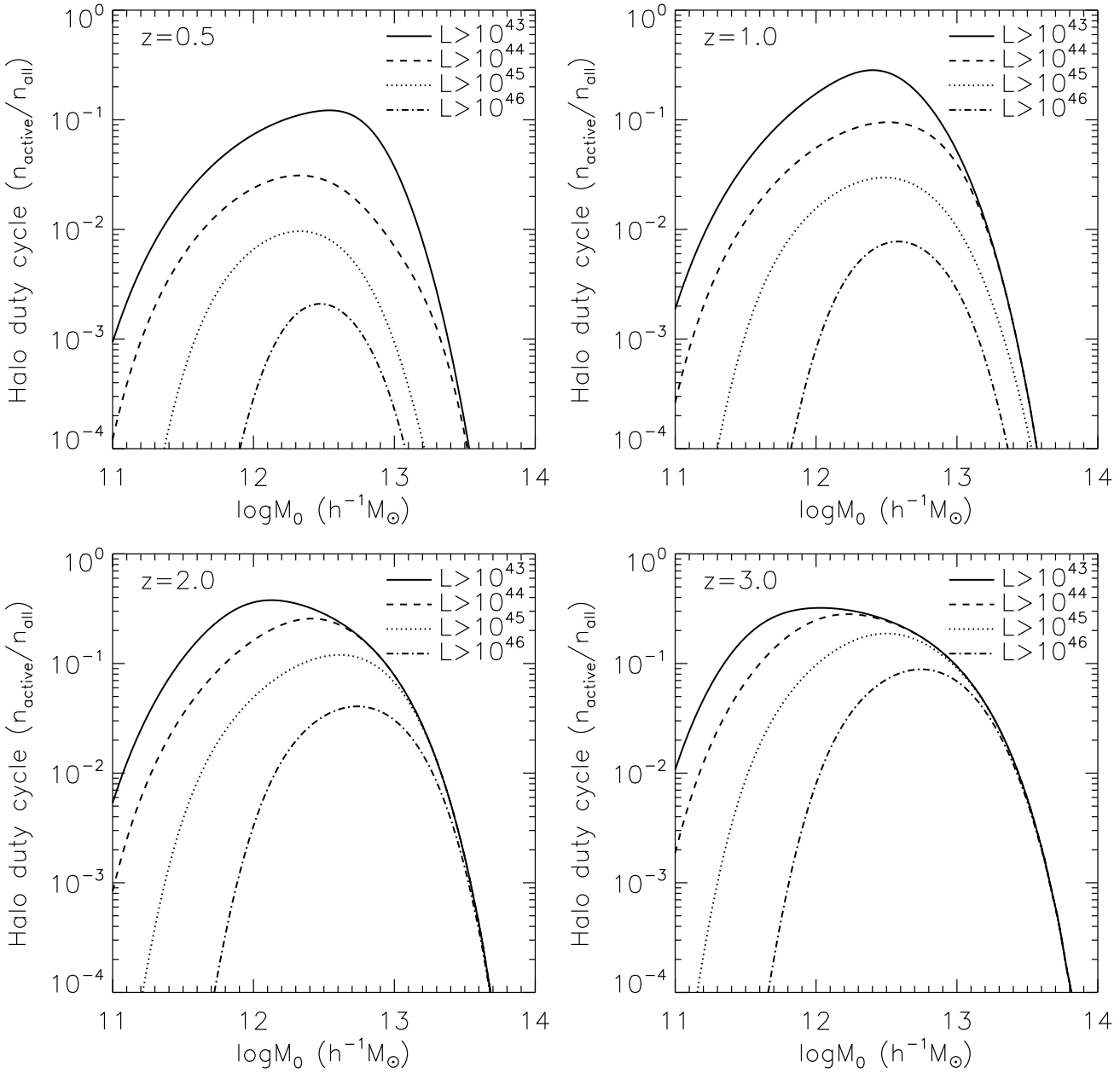

FIG. 8.— Halo duty cycles, i.e., fraction of halos hosting QSOs brighter than $L_{\rm min}$, as function of halo mass, computed using the halo mass function from Sheth et al. (2001) and the active halo mass function from Eqn. (21).

LF as an input (e.g., Yu & Tremaine 2002; Yu & Lu 2004, 2008; Marconi et al. 2004; Merloni 2004; Shankar et al. 2004, 2009b). Our simple framework, although semi-analytical in nature, accommodates a wide range of updated and new observations of QSO statistics; these new observations include quasar clustering and Eddington ratio distributions. Most of the early quasar models (e.g., Haiman & Loeb 1998; Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003; Volonteri et al. 2003) focused mainly on the luminosity function (partly because other observations were not available at that time), and most of them assumed simplified light curve models which were unable to reproduce the observed Eddington ratio distributions.

Hopkins et al. (2008) presented a merger-based quasar model that utilizes a variety of quasar observations for comparison, including the latest clustering measurements. Our model framework is different from theirs in two major aspects: 1) they estimated the major-merger rate from the combination of empirical halo occupation models of galaxies and merging timescale analysis, while we used directly the halo merger rate from simulations – both approaches have their own advantages and disadvantages; 2) the light curve in their model is extracted from their merger-event simulations (Hopkins et al. 2005), while we have adopted a parametrization of the light curve which is fit by observations. Their light curve model is clearly more physically-motivated than ours, yet it needs to be confirmed in future simulations with higher resolution and better understandings of BH accretion physics. On the other hand, the virtue of our framework is that it allows fast and easy estimations of model predictions and parameter adjustments to fit updated observations.

4.2. *Caveats in the Reference Model*

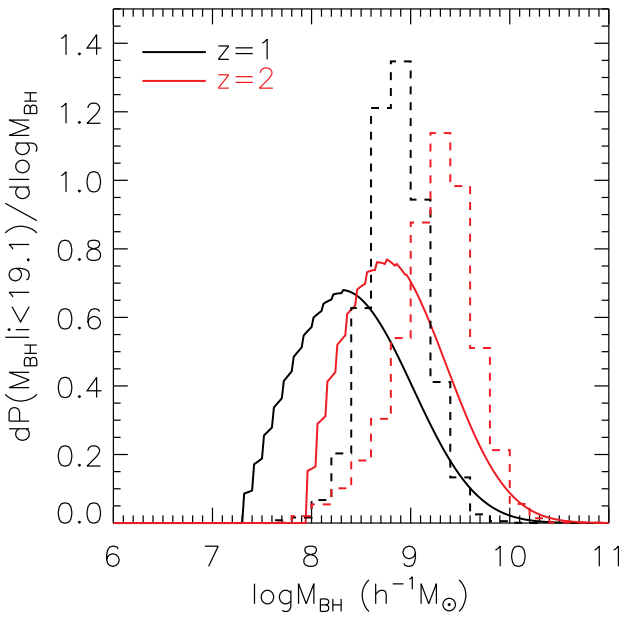

FIG. 9.— Distributions of BH masses in quasars at two redshifts. The solid lines are the predictions of our reference model for $i < 19.1$ (i.e., the flux limit in the main SDSS quasar catalog) quasars. The dashed lines show the distributions of *virial* BH masses from Shen et al. (2008a).

As already mentioned in §3, our fiducial model is not an actual χ^2 fit to the overall observations (LF, clustering, and Eddington ratio distributions) because of the ambiguity of assigning relative weights to individual observational data sets. Instead, we have experimented with varying the model parameters within reasonable ranges, to achieve a global "good" (as judged by eye) fit to observations. If consider only the LF as observational constraints, there are model degeneracies between the luminosity decaying rate α and the normalization of the mean $L_{\rm peak}-M_0$ relation C, i.e., if QSO luminosity decays more slowly, the typical host halos need to shift to more massive (and less abundant) halos in order not to overpredict the LF; likewise, if the scatter around the mean $L_{\rm peak}-M_0$ relation σ_L increases, we can reduce C to match the LF. However, these degeneracies are broken once the clustering observations are taken into account. A large scatter σ_L or slow varying light curve (small values of α) cannot fit the large bias at high redshift and the luminosity dependence of clustering. Our model is also more complicated than previous models (e.g., Haiman & Loeb 1998; Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002, 2003; Volonteri et al. 2003; Shankar et al. 2009b) in the sense that we have more parameters, which is required for a flexible enough framework to accommodate a variety of observations.

For our fiducial model we have used the halo merger rate as the proxy for the QSO-triggering rate. Alternatively, if we use the subhalo merger rate (the delayed version of halo merger rate; §1.4), it makes little difference below $z = 3$, but it further underestimates the LF at $z > 3$, as expected from Fig. 1. The bolometric LF data we used here has uncertainties both from measurements and bolometric corrections at the $\sim 20-30\%$ level (cf., Hopkins et al. 2007; Shankar et al. 2009b), not enough to reconcile the discrepancy at $z > 4.5$. There are several ways to modify our model to match observations of LF at $z > 4.5$: a) decrease the threshold $\xi_{\rm min}$ above which mergers can trigger QSO activity; b) modify the $L_{\rm peak}-M_0$ relation such that at fixed peak luminosity, QSOs shift to even smaller halos at $z > 4.5$, which can be achieved by including some higher order terms in the redshift evolution of C; and c) increase the scatter σ_L so that more abundant low mass halos can contribute to the LF by up-scattering. As an example of such a model, we modify the redshift evolution of the mean $L_{\rm peak}-M_0$ relation at high redshift such that $C \rightarrow C+\log[(\frac{1+z}{4.5})^{7/2}]$ for $z > 3.5$. With this additional term of evolution in the mean $L_{\rm peak}-M_0$ relation, the typical host halo shifts to lower masses ($M_0 \sim 4\times10^{11}\,h^{-1}M_\odot$ for $\langle \log L/{\rm ergs}^{-1}\rangle = 46$ at $z = 6$, compared to $M_0 \sim 10^{12}\,h^{-1}M_\odot$ in our fiducial model), and the resulting LF (dashed lines in Fig. 3) fit the observations well. This modification has little effects on the LF at $z < 3.5$, as well as other predicted QSO properties. An alternative parametrization of the redshift evolution in C is further discussed in §4.4.

However, there are no independent constraints on the mass of halos hosting quasars at $z > 4.5$ (such as those inferred from quasar clustering), and the merger scenario of quasar activity may be different at such high redshift. Hence we do not attempt to fully resolve this issue in this paper.

There is also slight tension between quasar clustering and LF at $z \sim 4$ in our model, as already noted by several studies (White et al. 2008; Wyithe & Loeb 2009; Shankar et al. 2009a). On the one hand, we need smaller halo masses to account for quasar abundance; we also need larger halos to account for the strong clustering on the other. To fully resolve this issue we need better understanding of the halo bias, *and reliable fitting formulae for it*, derived from simulations for the relevant mass and redshift ranges[8], as well as better measurements of quasar clustering at high redshift with future larger samples. Nevertheless, our fiducial model is still consistent with both LF and clustering observations within the errors.

Our model also underpredicts the LF at the low luminosity end ($L < 10^{45}\,{\rm ergs}^{-1}$) at $z < 0.5$. This is somewhat expected, since our model does not include contributions from AGNs triggered by mechanisms other than a major merger. The fuel budget needed to feed a low luminosity AGN ($L < 10^{45}\,{\rm ergs}^{-1}$) is much less stringent than that for bright quasars. Therefore secular processes (i.e., gas inflows driven by bars, tidal encounters, stochastic accretion, as well as minor mergers), while not as violent and efficient as major mergers, provide viable means to fuel AGNs at low activity levels. At $z < 0.5$, there are evidence that some low luminosity AGN ($L \sim 10^{43-44}\,{\rm ergs}^{-1}$, powered by intermediate-mass BHs $M_\bullet \sim 10^6 M_\odot$) hosts have no classical bulges[9], which indicates that secular processes are responsible for triggering these low-luminosity AGN activity and BH growth (e.g., Greene et al. 2008, and references therein). There are also

[8] For instance, it has been suggested that in addition to mass, halo clustering also depends on concentration, assembly history and recent merger activity (the so-called assembly bias, e.g., Gao et al. 2005; Wechsler et al. 2006; Wetzel et al. 2007). If recently merged halos (as quasar hosts) have larger bias than average for the same halo mass, then it may reconcile the slight tension between reproducing both the LF and quasar clustering at high redshift (e.g., Wyithe & Loeb 2009), although current estimate of this enhancement is only on the level of $\sim 10\%$ (e.g., Wetzel et al. 2007).

[9] Classical bulges (including ellipticals and bulges in early-type disk galaxies) are presumably formed via mergers and they follow the fundamental plane of elliptical galaxies (e.g., Bender et al. 1992). On the other hand, secular processes can build up the so-called *pseudobulges*, which have distinct structural properties from classical bulges (Kormendy & Kennicutt 2004). While pseudobulges can host central BHs, there are some evidence that the bulge-BH mass relations for pseudobulge systems are offset from that for classical bulge systems (e.g., Hu 2008; Greene et al. 2008), indicating that secular processes are less efficient in building BHs.

observational implications that the bulk of BH growth has shifted from the most massive BHs ($M_\bullet > 10^8 M_\odot$) at high redshift ($z \lesssim 2$) to low mass BHs ($M_\bullet < 10^8 M_\odot$) locally (e.g., McLure & Dunlop 2004; Heckman et al. 2004), along with the cosmic downsizing in luminosity function evolution (e.g., Steffen et al. 2003; Ueda et al. 2003; Hasinger et al. 2005; Hopkins et al. 2007; Bongiorno et al. 2007). The typical transition from merger-driven BH growth to secularly-driven BH growth likely occurs around BH mass $\sim 10^7 M_\odot$ (e.g., Hopkins & Hernquist 2009), corresponding to $L \sim 10^{45}$ ergs^{-1} at the Eddington limit. Since the specific merger rate increases towards higher redshift, while the rate of secular processes is almost constant with time and these processes are relatively slow and inefficient for BH growth, it is conceivable that secular processes will only become important at late times ($z < 0.5$ for instance) in building up the low mass end of the SMBH population. The implementation of secularly-driven AGN activity in our model will be presented in future work.

Finally, we mention that our model predicts a turnover in the LF at $L \lesssim 10^{44}$ ergs^{-1} at $z \gtrsim 3$. This simply reflects the lower mass cutoff at $M_{\rm min} = 3 \times 10^{11} h^{-1} M_\odot$ in our model. Future deeper surveys for low luminosity AGNs at $z \gtrsim 3$ are necessary to probe this luminosity regime, and to impose constraints on how efficiently SMBHs can form in low mass halos.

4.3. *The Effects of BH Coalescence on the BHMF*

The BHMFs discussed in §3 and Fig. 7 are the mass functions from accretion only, i.e., we have neglected the effects of BH coalescence. As discussed in §1.1, BH coalescence will redistribute the BH mass function but does not change the total BH mass density (neglecting mass loss via gravitational radiation) since essentially all the mass ended up in BHs were accreted. There are two routes in which the coalescence with pre-existing BHs might become important within our model framework[10]. First, both galaxies in the merging pair of halos are elliptical, i.e., they already experienced a QSO phase in the past and formed massive nuclear BHs. In this case the current merger event will be a dry merger and form a SMBH binary without triggering a new QSO. Second, during the major merger, one galaxy is elliptical and the other one is spiral, in which case a QSO will be triggered and the mass of the old BH of the previous elliptical will add to the new BH system.

A complete exploration of these routes and their consequences (including the effects of BH ejection, mass loss through gravitational radiation, etc.) can be better achieved with Monte Carlo realizations of halo merger trees and our model prescriptions for QSO triggering and BH growth, which we plan to investigate in a future paper. Here we can approximately estimate the maximal impact of BH coalescence on the BHMFs in the two cases, assuming that BH coalescence always occurs within a Hubble time and all the mass in the pre-existing BHs is added to the final BH, i.e., no BH ejection or mass loss via gravitational radiation. We focus on the high-mass end ($M_\bullet \gtrsim 10^8 M_\odot$) of the local BHMF since our model prediction is incomplete at the lower-mass end.

Although the *dry merger* case is not subject to the major merger condition for QSO-triggering, we still restrict to $\xi \geq \xi_{\rm min}$ here because in minor mergers: 1) it will take too long for the two pre-existing BHs to become a close binary, and 2) the mass increment due to BH coalescence is insignificant. Observational determination of the major dry merger rate is difficult, and the current best estimate is: on average, present-day spheroidal galaxies with $M_V < -20.5$ (corresponding to $M_\bullet \gtrsim 10^8 M_\odot$) have undergone $0.5-2$ major dyr mergers since $z \sim 0.7$ (Bell et al. 2006). If we assume all the $M_\bullet \gtrsim 10^8 M_\odot$ BHs undergo one $1:1$ dry merger after the QSO phase, the local BHMF will redistribute as the yellow line in Fig. 7. In this case the abundance of the most massive ($M_\bullet >$ a few $\times 10^9 M_\odot$) BHs is enhanced by up to a factor of ~ 2 at $M_\bullet = 10^{10} h^{-1} M_\odot$.

In the *half-dry major merger* case, the fraction of the pre-existing BH mass to the final BH mass is $(1+\xi')^{-5/3} \sim 0.07-0.7$, where $1/4 \leq \xi' \leq 4$ is the major merger mass ratio (in our reference model) between the two halos. This is because the final BH mass after a QSO phase scales as the $5/3$ power to the halo mass in our model. Averaging over possible values of ξ', the fractional increment due to the pre-existing BH is $\sim 30\%$. If all QSO-triggering mergers are this kind of *half-dry* event, the predicted $z = 0$ BHMF (red line in Fig. 7) will redistribute from lower mass to higher mass (the green line). The enhancement at the high-mass end of the local BHMF is comparable to the dry major merger case. In practice QSO-triggering mergers can occur between two spirals, hence the actual correction due to pre-existing BHs in *half-dry* mergers should be smaller.

Combining these two cases we conclude that the impact of BH coalescence on the BHMF is probably insignificant compared with other uncertainties and systematics in current observations and our model framework. Similar conclusions were also achieved in several independent work (Volonteri et al. 2003; Yu & Lu 2008; Shankar et al. 2009b) albeit with difference in details.

4.4. *Implications for BH Scaling Relations*

In our formalism there is a generic scaling relation between the relic BH mass and halo mass, which has the same slope and scatter as the $L_{\rm peak} - M_0$ relation (from Eqns. 17 and 29):

$$\frac{M_{\bullet,\rm relic}}{10^8 h^{-1} M_\odot} \approx 0.6(1+z)^{\beta_1} \left(\frac{M_0}{10^{12} h^{-1} M_\odot} \right)^{5/3}. \quad (35)$$

The local $M_\bullet - M_0$ relation for dormant BHs reported in Ferrarese (2002) and Baes et al. (2003) has a slope in the range $\sim 1.3-1.8$, and a normalization lower by a factor of $\sim 3-5$ than our predictions. This is probably due to the fact that we neglected continued growth of halos by minor mergers and diffuse matter accretion since the major merger event, if most of the local massive BHs were assembled at $z \gtrsim 1$ (as our model predicts). The fact that we did not impose a cutoff in the LC (24) may also lead to overly massive BHs. We note that although the normalization (and perhaps slope as well) of the local $M_\bullet - M_0$ relation may depend on galaxy morphological type (e.g., Zasov et al. 2004; Courteau et al. 2007; Ho 2007), early-type galaxies (S0 and ellipticals) appear to occupy the upper envelope in the $M_\bullet - M_0$ relation. Our model scaling relations are only valid for early type galaxies which are presumably merger remnants.

However, it should be pointed out that the BH-halo scaling relation relies on the assumed $L_{\rm peak} - M_0$ relation. The simple prescription for its evolution in our reference model already shows some difficulties in reproducing the QSO LF at $z \gtrsim 4.5$

[10] In contrary to the two cases discussed below, we assume that a major merger between two spirals will lead to negligible BH mass contribution from the pre-existing BHs, since in our model setting, spiral galaxy has not yet experienced a major merger and hence significant BH growth.

(Fig. 3; see §4.2). Hence our fiducial model is not very appropriate for predicting the redshift evolution of the BH-halo scaling relation. To make this point more clear, let us consider a more physically-motivated prescription for the $L_{\rm peak}-M_0$ relation in which $L\propto V_{\rm vir}^5$, where $V_{\rm vir}$ is the halo virial velocity (Eqn. A5), and the normalization of the $L_{\rm peak}-M_0$ relation evolves as (Wyithe & Loeb 2002, 2003):

$$C(z)=C(z=0)+\frac{5}{2}\log(1+z)+\frac{5}{6}\log\left[\frac{\Delta_{\rm vir}(z)}{\Delta_{\rm vir}(0)}\right], \qquad (36)$$

e.g., the evolution in $C(z)$ is more rapid than our fiducial setting. With this new implementation for the $L_{\rm peak}-M_0$ relation, we found that a good global fit can be achieved with the following parameter adjustments (other parameters are the same as in our reference model): $M_{\rm quench}=3\times10^{12}\,h^{-1}M_\odot$, $C(z=0)=\log(0.8\times10^{45})-12\gamma$ and $\sigma_L=0.4(1+z)^{-1/2}$. Since the redshift evolution in the normalization $C(z)$ is now faster, the starting value $C(0)$ is reduced; consequently the exponential upper cut of halo mass, $M_{\rm quench}$, increases in order to account for QSO counts at low redshift. The scatter in the $L_{\rm peak}-M_0$ relation needs a redshift evolution to achieve adequate fits for both clustering and LF over a wide redshift range. The predicted LF is shown as dotted lines in Fig. 3, which does a much better job at $z\gtrsim4.5$ than our fiducial model. Other predicted properties are slightly degraded (but still are reasonably good fits to observations) than those predicated by our fiducial model.

This new $L_{\rm peak}-M_0$ relation, together with the approximation that the halo virial velocity $V_{\rm vir}\approx v_c$, the galaxy circular velocity, and the assumption that the local $v_c-\sigma$ relation (Ferrarese 2002) does not evolve, result in a constant $M_\bullet-\sigma$ relation (Wyithe & Loeb 2003). Neglecting the scatter, the local mean $v_c-\sigma$ relation in Ferrarese (2002) is:

$$\log\left(\frac{v_c}{\rm km\,s^{-1}}\right)=0.84\log\left(\frac{\sigma}{\rm km\,s^{-1}}\right)+0.55\,. \qquad (37)$$

Thus the new $L_{\rm peak}-M_0$ relation predicts a constant $M_\bullet-\sigma$ relation (e.g., Eqns. 17, 26, 29, 36, and A6):

$$\frac{M_\bullet}{10^8\,M_\odot}=0.9\sim5.0\times\left(\frac{\sigma}{200\,\rm km\,s^{-1}}\right)^{4.2}, \qquad (38)$$

with a slope and normalization (bounded by $M_{\bullet,\rm peak}$ and $M_{\bullet,\rm relic}$) consistent with local estimates (Gebhardt et al. 2000; Ferrarese & Merritt 2000; Tremaine et al. 2002; Lauer et al. 2007). This consistency, however, is built on a couple of assumptions and approximations, and neglecting successive evolution after the self-regulation of BH and bulge growth. Given all these complications, it is beyond the scope of the current study to fully settle this issue. We simply remind the reader that a non-evolving $M_\bullet-\sigma$ relation is generally allowed within our framework.

5. CONCLUSIONS

We have developed a general cosmological framework for the growth and cosmic evolution of SMBHs in the hierarchical merging scenario. Assuming that QSO activity is triggered by major mergers of host halos, and that the resulting light curve follows a universal form with its peak luminosity correlated with the (post)merger halo mass, we model the QSO LF and SMBH growth self-consistently across cosmic time. We tested our model against a variety of observations of SMBH statistics: the QSO luminosity function, quasar clustering, quasar/AGN BH mass and Eddington ratio distributions. A global good fit is achieved with reasonable parameters. We summarize our model specifics as follows:

- The QSO-triggering rate is determined by the $\xi\geq0.25$ halo merger rate with exponential cutoffs at both the low and high halo mass ends $M_{\rm min}=3\times10^{11}\,h^{-1}M_\odot$, $M_{\rm max}(z)=10^{12}(1+z)^{3/2}\,h^{-1}M_\odot$.

- The universal light curve follows an initial exponential Salpeter growth with constant Eddington ratio $\lambda_0=3$ for a few e-folding times to reach the peak luminosity $L_{\rm peak}$, which is correlated with the (post)merger halo mass as $\langle L_{\rm peak}\rangle=6\times10^{45}(1+z)^{1/3}(M_0/10^{12}\,h^{-1}M_\odot)^{5/3}\,\rm ergs^{-1}$, with a log-normal scatter $\sigma_L=0.28$ dex. It then decays as a power-law with slope $\alpha=2.5$.

Our simple model successfully reproduces the LF, quasar clustering, and Eddington ratio distributions of quasars and AGNs at $0.5<z<4.5$, supporting the hypothesis that QSO activity is linked to major merger events within this redshift range. However, there are still many unsettled issues. Below we outline several possible improvements of our simple model, which will be addressed in future work.

Our model under-predicts the LF at the faint luminosity end $L<10^{45}\,\rm ergs^{-1}$ at $z<0.5$, which is linked to the growth of the less massive $\lesssim10^7\,M_\odot$ SMBHs. This is indicative of a population of low luminosity AGNs triggered by mechanisms other than major mergers at the low redshift universe – either by minor mergers or secular processes. We need to incorporate this ingredient in our SMBH model, in order to match the local BHMF at the low mass end ($M_\bullet\lesssim10^7\,M_\odot$).

In our modeling we have neglected the possibilities of a closely-following second major merger event and the triggering of two simultaneous QSOs during a single major merger event. Therefore our model does not include more than one QSOs within a single halo. We will use Monte-Carlo realizations of halo merger trees to assess the probability of such rare occurrences and see if they can account for the small ($\lesssim0.1\%$) binary/multiple quasar fraction observed (Hennawi et al. 2006, 2009; Myers et al. 2008).

Our model can be improved to include the radio loudness of QSOs as well. If radio loudness requires both a massive host halo (to provide the hot IGM) and a massive SMBH (to launch the kinetic jet), we can statistically populate radio-loud QSOs in halos within our model framework. It can be tested against the clustering of radio-loud quasars (e.g., Shen et al. 2009) and the radio-loud fraction as function of luminosity and redshift (e.g., Jiang et al. 2007, and references therein).

We thank the anonymous referee, Francesco Shankar, Michael Strauss, Scott Tremaine, and Martin White for constructive comments that have greatly improved the manuscript. We are grateful to Francesco Shankar for pointing out an error in the merger rate equation (4) in an earlier version of the manuscript. This work was supported by NSF grant AST-0707266.

APPENDIX

DARK MATTER HALOS

All dark matter halos are assumed to have a spherical NFW profile (Navarro et al. 1997):

$$\rho_{\rm NFW}(r) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2} \ , \tag{A1}$$

where r_s and ρ_s are the characteristic scale and the density at this scale.

The virial mass and virial radius are related by (Bullock et al. 2001):

$$M_{\rm vir} \equiv \frac{4\pi}{3}\Delta_{\rm vir}\rho_u r_{\rm vir}^3 \ , \tag{A2}$$

where $\Delta_{\rm vir}(z) \approx (18\pi^2 + 82x - 39x^2)/\Omega(z)$ with $x \equiv \Omega(z) - 1$ (Bryan & Norman 1998) is the spherical overdensity relative to the background *matter* density ρ_u and $\Omega(z) = [1+\Omega_\Lambda(1+z)^{-3}/\Omega_0]^{-1}$. Note there is the slight difference in defining the virial radius in Bullock et al. (2001) and Navarro et al. (1997) where the latter uses r_{200} (the radius corresponding to a spherical overdensity 200 times the *critical* density) to define the virial radius. Both definitions of virial radius are frequently used in studies on the galaxy merger time scale within merged dark matter halos: Jiang et al. (2008), Stewart et al. (2008) used the former definition, while Boylan-Kolchin et al. (2008), Wetzel et al. (2009) used the latter.

The enclosed mass within an NFW profile truncated at r_o is

$$M(r_o) = \int_0^{r_o} dr 4\pi r^2 \rho_{\rm NFW}(r) = 4\pi\rho_s r_s^3 \left[\ln(1+c) - \frac{c}{1+c}\right] \ , \tag{A3}$$

where $c \equiv r_o/r_s$. Therefore we have

$$M_{\rm vir} = M(r_{\rm vir}) = 4\pi\rho_s r_s^3 \left[\ln(1+c_{\rm vir}) - \frac{c_{\rm vir}}{1+c_{\rm vir}}\right] \tag{A4}$$

where $c_{\rm vir} \equiv r_{\rm vir}/r_s$ is the usual definition of the concentration parameter. The mean relation between $M_{\rm vir}$ and $c_{\rm vir}$ is given in Bullock et al. (2001).

The virial velocity (usually defined as the circular velocity at the virial radius) $V_{\rm vir}$, and the maximum circular velocity at $r_{\rm max} \approx 2.16 r_s$ are (Bullock et al. 2001):

$$V_{\rm vir}^2 \equiv V_c^2(r_{\rm vir}) = \frac{GM_{\rm vir}}{r_{\rm vir}} \ , \qquad \frac{V_{\rm max}^2}{V_{\rm vir}^2} \approx \frac{0.216 c_{\rm vir}}{\ln(1+c_{\rm vir}) - c_{\rm vir}/(1+c_{\rm vir})} \ . \tag{A5}$$

This implies that the relation between virial mass $M_{\rm vir}$ and virial velocity $V_{\rm vir}$ is:

$$M_{\rm vir}(z) = \left[\frac{4\pi}{3}\Delta_{\rm vir}(z)\rho_0\right]^{-1/2} (1+z)^{-3/2} G^{-3/2} V_{\rm vir}^3 = 1.37\times10^{12}\Omega_0^{-1/2}h^{-1}M_\odot \left(\frac{V_{\rm vir}}{200\,{\rm km\,s^{-1}}}\right)^3 (1+z)^{-3/2}\left[\frac{\Delta_{\rm vir}(z)}{\Delta_{\rm vir}(0)}\right]^{-1/2} \ , \tag{A6}$$

where $\rho_0 = 2.78\times10^{11}\Omega_0 h^2 M_\odot {\rm Mpc}^{-3}$ is the $z = 0$ mean matter density.

The DM halo dynamical time $\tau_{\rm dyn}$ is usually defined as $r_{\rm vir}/V_{\rm vir}$:

$$\tau_{\rm dyn} \equiv \frac{r_{\rm vir}}{V_{\rm vir}} = 1.4\times10^{10}\,{\rm yr}\times[\Omega_0 h^2 \Delta_{\rm vir}(1+z)^3]^{-1/2} \ . \tag{A7}$$

For simplicity we have neglected the difference between $M_{\rm vir}$ and the *friends-of-friends* mass $M_{\rm fof}$ (with a link length $b = 0.2$) throughout the paper. But we give an approximate conversion formula below for completeness:

$$\frac{M_{\rm fof}}{M_{\rm vir}} = \frac{\ln(1+c_{\rm fof}) - c_{\rm fof}/(1+c_{\rm fof})}{\ln(1+c_{\rm vir}) - c_{\rm vir}/(1+c_{\rm vir})} \ , \tag{A8}$$

where $c_{\rm fof} = r_{\rm fof}/r_s$ can be solved via the following equation:

$$c_{\rm fof}(1+c_{\rm fof})^2 = \frac{2\pi b^3\Delta_{\rm vir}}{9}\frac{c_{\rm vir}^3}{\ln(1+c_{\rm vir}) - c_{\rm vir}/(1+c_{\rm vir})} \ , \tag{A9}$$

where we have used the fact that the density at $r_{\rm fof}$ is $\rho_{\rm NFW}(r_{\rm fof}) \approx 3\rho_u/(2\pi b^3)$.

REFERENCES

Adelberger, K. L., & Steidel, C. C. 2005, ApJ, 630, 50
Babić, A., Miller, L., Jarvis, M. J., Turner, T. J., Alexander, D. M., & Croom, S. M. 2007, A&A, 474, 755
Baes, M., Buyle, P., Hau, G. K. T., & Dejonghe, H. 2003, MNRAS, 341, L44
Bahcall, J. N., Kirhakos, S., Saxe, D. H., & Schneider, D. P. 1997, ApJ, 479, 642
Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15
Barger, A. J., Cowie, L. L., Mushotzky, R. F., Yang, Y., Wang, W.-H., Steffen, A. T., & Capak, P. 2005, AJ, 129, 578
Begelman, M. C. 2002, ApJ, 568, L97
Bell, E. F., et al. 2006, ApJ, 640, 241
Bender, R., Burstein, D., & Faber, S. M. 1992, ApJ, 399, 462
Bennert, N., Canalizo, G., Jungwiert, B., Stockton, A., Schweizer, F., Peng, C. Y., & Lacy, M. 2008, ApJ, 677, 846
Binney, J., & Tremaine, S. 1987, Galactic Dynamics (Princeton, NJ, Princeton University Press, 1987, 747 p.)
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Bond, J. R., & Myers, S. T. 1996, ApJS, 103, 1
Bongiorno, A., et al. 2007, A&A, 472, 443
Bonoli, S., Marulli, F., Springel, V., White, S. D. M., Branchini, E., & Moscardini, L. 2009, MNRAS, 396, 423
Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2008, MNRAS, 383, 93
Bryan, G. L., & Norman, M. L. 1998, ApJ, 495, 80
Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R., & Dekel, A. 2001, MNRAS, 321, 559
Canalizo, G., & Stockton, A. 2001, ApJ, 555, 719
Carlberg, R. G. 1990, ApJ, 350, 505
Chandrasekhar, S. 1943, ApJ, 97, 255
Cole, S., & Kaiser, N. 1989, MNRAS, 237, 1127
Courteau, S., McDonald, M., Widrow, L. M., & Holtzman, J. 2007, ApJ, 655, L21
Croom, S. M., et al. 2005, MNRAS, 356, 415
Croom, S. M., Smith, R. J., Boyle, B. J., Shanks, T., Miller, L., Outram, P. J., & Loaring, N. S. 2004, MNRAS, 349, 1397
Croton, D. J. 2009, MNRAS, 394, 1109
Croton, D. J., et al. 2006, MNRAS, 365, 11
da Ângela, J., et al. 2008, MNRAS, 383, 565
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604
Eisenstein, D. J., & Hu, W. 1999, ApJ, 511, 5
Fakhouri, O., & Ma, C.-P. 2008, MNRAS, 386, 577
Fan, X., et al. 2004, AJ, 128, 515
—. 2001, AJ, 121, 54
Ferrarese, L. 2002, ApJ, 578, 90
Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9
Fisher, K. B., Bahcall, J. N., Kirhakos, S., & Schneider, D. P. 1996, ApJ, 468, 469
Fontanot, F., Cristiani, S., Monaco, P., Nonino, M., Vanzella, E., Brandt, W. N., Grazian, A., & Mao, J. 2007, A&A, 461, 39
Francke, H., et al. 2008, ApJ, 673, L13
Gao, L., Springel, V., & White, S. D. M. 2005, MNRAS, 363, L66
Gavignaud, I., et al. 2008, A&A, 492, 637
Gebhardt, K., et al. 2000, ApJ, 539, L13
Graham, A. W. 2008, ApJ, 680, 143
Graham, A. W., Erwin, P., Caon, N., & Trujillo, I. 2001, ApJ, 563, L11
Graham, A. W., & Li, I.-h. 2009, ApJ, 698, 812
Granato, G. L., De Zotti, G., Silva, L., Bressan, A., & Danese, L. 2004, ApJ, 600, 580
Greene, J. E., Ho, L. C., & Barth, A. J. 2008, ApJ, 688, 159
Gunn, J. E., & Gott, J. R. I. 1972, ApJ, 176, 1
Haiman, Z., & Hui, L. 2001, ApJ, 547, 27
Haiman, Z., & Loeb, A. 1998, ApJ, 503, 505
Hasinger, G., Miyaji, T., & Schmidt, M. 2005, A&A, 441, 417
Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., Tremonti, C., & White, S. D. M. 2004, ApJ, 613, 109
Hennawi, J. F., et al. 2006, AJ, 131, 1
—. 2009, ArXiv e-prints
Hernquist, L. 1989, Nature, 340, 687
Ho, L. C. 2007, ApJ, 668, 94
Hopkins, P. F., & Hernquist, L. 2009, ApJ, 694, 599
Hopkins, P. F., Hernquist, L., Cox, T. J., & Kereš, D. 2008, ApJS, 175, 356
Hopkins, P. F., Hernquist, L., Martini, P., Cox, T. J., Robertson, B., Di Matteo, T., & Springel, V. 2005, ApJ, 625, L71
Hopkins, P. F., Narayan, R., & Hernquist, L. 2006, ApJ, 643, 641
Hopkins, P. F., Richards, G. T., & Hernquist, L. 2007, ApJ, 654, 731
Hu, J. 2008, MNRAS, 386, 2242
Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, MNRAS, 321, 372
Jiang, C. Y., Jing, Y. P., Faltenbacher, A., Lin, W. P., & Li, C. 2008, ApJ, 675, 1095
Jiang, L., et al. 2006, AJ, 131, 2788
Jiang, L., Fan, X., Ivezić, Ž., Richards, G. T., Schneider, D. P., Strauss, M. A., & Kelly, B. C. 2007, ApJ, 656, 680
Kaiser, N. 1984, ApJ, 284, L9
Kauffmann, G., & Haehnelt, M. 2000, MNRAS, 311, 576
King, A. 2003, ApJ, 596, L27
Kollmeier, J. A., et al. 2006, ApJ, 648, 128
Kormendy, J., & Kennicutt, Jr., R. C. 2004, ARA&A, 42, 603
Kormendy, J., & Richstone, D. 1995, ARA&A, 33, 581
Lacey, C., & Cole, S. 1993, MNRAS, 262, 627
Lapi, A., Shankar, F., Mao, J., Granato, G. L., Silva, L., De Zotti, G., & Danese, L. 2006, ApJ, 650, 42
Lauer, T. R., et al. 2007, ApJ, 662, 808
Lidz, A., Hopkins, P. F., Cox, T. J., Hernquist, L., & Robertson, B. 2006, ApJ, 641, 41
Lynden-Bell, D. 1969, Nature, 223, 690
Magorrian, J., et al. 1998, AJ, 115, 2285
Marconi, A., & Hunt, L. K. 2003, ApJ, 589, L21
Marconi, A., Risaliti, G., Gilli, R., Hunt, L. K., Maiolino, R., & Salvati, M. 2004, MNRAS, 351, 169
Martini, P., & Weinberg, D. H. 2001, ApJ, 547, 12
McLure, R. J., & Dunlop, J. S. 2004, MNRAS, 352, 1390
Merloni, A. 2004, MNRAS, 353, 1035
Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347
Monaco, P., Fontanot, F., & Taffoni, G. 2007, MNRAS, 375, 1189
Myers, A. D., Brunner, R. J., Nichol, R. C., Richards, G. T., Schneider, D. P., & Bahcall, N. A. 2007a, ApJ, 658, 85
Myers, A. D., Brunner, R. J., Richards, G. T., Nichol, R. C., Schneider, D. P., & Bahcall, N. A. 2007b, ApJ, 658, 99
Myers, A. D., et al. 2006, ApJ, 638, 622
Myers, A. D., Richards, G. T., Brunner, R. J., Schneider, D. P., Strand, N. E., Hall, P. B., Blomquist, J. A., & York, D. G. 2008, ApJ, 678, 635
Narayan, R., & Yi, I. 1995, ApJ, 452, 710
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Padmanabhan, N., White, M., Norberg, P., & Porciani, C. 2008, ArXiv e-prints
Porciani, C., Magliocchetti, M., & Norberg, P. 2004, MNRAS, 355, 1010
Porciani, C., & Norberg, P. 2006, MNRAS, 371, 1824
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Richards, G. T., et al. 2005, MNRAS, 360, 839
—. 2006, AJ, 131, 2766
Richstone, D., et al. 1998, Nature, 395, A14
Ross, N. P., et al. 2009, ApJ, 697, 1634
Salpeter, E. E. 1964, ApJ, 140, 796
Salucci, P., Szuszkiewicz, E., Monaco, P., & Danese, L. 1999, MNRAS, 307, 637
Sanders, D. B., & Mirabel, I. F. 1996, ARA&A, 34, 749
Schmidt, M., & Green, R. F. 1983, ApJ, 269, 352
Schneider, D. P., et al. 2007, AJ, 134, 102
Serber, W., Bahcall, N., Ménard, B., & Richards, G. 2006, ApJ, 643, 68
Shankar, F., Crocce, M., Miralda-Escude', J., Fosalba, P., & Weinberg, D. H. 2009a, ArXiv e-prints
Shankar, F., Salucci, P., Granato, G. L., De Zotti, G., & Danese, L. 2004, MNRAS, 354, 1020
Shankar, F., Weinberg, D. H., & Miralda-Escudé, J. 2009b, ApJ, 690, 20
Shaver, P. A. 1984, A&A, 136, L9
Shen, Y., Greene, J. E., Strauss, M. A., Richards, G. T., & Schneider, D. P. 2008a, ApJ, 680, 169
Shen, Y., Strauss, M. A., Hall, P. B., Schneider, D. P., York, D. G., & Bahcall, N. A. 2008b, ApJ, 677, 858
Shen, Y., et al. 2007, AJ, 133, 2222
—. 2009, ApJ, 697, 1656
Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1
Sheth, R. K., & Tormen, G. 1999, MNRAS, 308, 119
Silk, J., & Rees, M. J. 1998, A&A, 331, L1
Silverman, J. D., et al. 2005, ApJ, 624, 630
Small, T. A., & Blandford, R. D. 1992, MNRAS, 259, 725
Soltan, A. 1982, MNRAS, 200, 115
Springel, V., Di Matteo, T., & Hernquist, L. 2005a, MNRAS, 361, 776
Springel, V., et al. 2005b, Nature, 435, 629
Steffen, A. T., Barger, A. J., Cowie, L. L., Mushotzky, R. F., & Yang, Y. 2003, ApJ, 596, L23
Stewart, K. R., Bullock, J. S., Barton, E. J., & Wechsler, R. H. 2008, ArXiv e-prints

Strand, N. E., Brunner, R. J., & Myers, A. D. 2008, ApJ, 688, 180
Taffoni, G., Mayer, L., Colpi, M., & Governato, F. 2003, MNRAS, 341, 434
Thacker, R. J., Scannapieco, E., Couchman, H. M. P., & Richardson, M. 2009, ApJ, 693, 552
Tinker, J., Kravtsov, A. V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., Gottlöber, S., & Holz, D. E. 2008, ApJ, 688, 709
Tremaine, S., et al. 2002, ApJ, 574, 740
Tundo, E., Bernardi, M., Hyde, J. B., Sheth, R. K., & Pizzella, A. 2007, ApJ, 663, 53
Ueda, Y., Akiyama, M., Ohta, K., & Miyaji, T. 2003, ApJ, 598, 886
Volonteri, M., Haardt, F., & Madau, P. 2003, ApJ, 582, 559
Warren, M. S., Abazajian, K., Holz, D. E., & Teodoro, L. 2006, ApJ, 646, 881
Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., & Allgood, B. 2006, ApJ, 652, 71
Wetzel, A. R., Cohn, J. D., & White, M. 2009, MNRAS, 395, 1376
Wetzel, A. R., Cohn, J. D., White, M., Holz, D. E., & Warren, M. S. 2007, ApJ, 656, 139
White, M. 2002, ApJS, 143, 241
White, M., Martini, P., & Cohn, J. D. 2008, MNRAS, 390, 1179
Wolf, C., Wisotzki, L., Borch, A., Dye, S., Kleinheinrich, M., & Meisenheimer, K. 2003, A&A, 408, 499
Woo, J.-H., & Urry, C. M. 2002, ApJ, 579, 530
Wyithe, J. S. B., & Loeb, A. 2002, ApJ, 581, 886
—. 2003, ApJ, 595, 614
—. 2009, MNRAS, 395, 1607
York, D. G., et al. 2000, AJ, 120, 1579
Yu, Q., & Lu, Y. 2004, ApJ, 602, 603
—. 2008, ApJ, 689, 732
Yu, Q., & Tremaine, S. 2002, MNRAS, 335, 965
Zasov, A. V., Khoperskov, A. V., & Tyurina, N. V. 2004, Astronomy Letters, 30, 593
Zel'dovich, Y. B., & Novikov, I. D. 1964, Dokl. Akad. Nauk SSSR, 158, 811
Zhang, J., & Hui, L. 2006, ApJ, 641, 641
Zhang, J., Ma, C.-P., & Fakhouri, O. 2008, MNRAS, 387, L13